UNIVERSIDAD POLITÉCNICA DE MADRID

POLITÉCNICA

ETS Ingenieros Informáticos

Métodos y Técnicas para la Evaluación del aprendizaje de Ontologías de Dominio

TESIS FIN DE MÁSTER

Autor:
David Chaves Fraga

Supervisor:
Oscar Corcho

Máster Universitario en Inteligencia Artificial

Julio 2016

Agradecimientos

Este trabajo no habría sido posible sin la ayuda de mi tutor, Oscar Corcho, al que le agradezco el tiempo dedicado y sus consejos en todos los aspectos del trabajo que, además, me han servido para introducirme e ilusionarme con el mundo de la investigación. También quiero agradecer a Carlos su paciencia al explicarme los sistemas de aprendizaje de *DrInventor* y a José Luis sus sugerencias y recomendaciones tanto en el estado del arte como en el diseño de los experimentos.

Al resto de personas que forman el OEG, por su acogida en el grupo, haciéndome sentir muy cómodo en el laboratorio a lo largo de los cinco meses que he estado desarrollando este trabajo. A mi familia, ya que sin su apoyo, aunque desde la lejanía, la realización de este máster habría sido casi imposible. A mis amigos de siempre, con los que he tenido la suerte de reencontrarme en Madrid. A Ana, por saber escuchar. Y, para terminar, especialmente a Alba, por su compañía y apoyo durante este último año.

Resumen

Desde su invención, la creación de ontologías se ha llevado a cabo, casi siempre, de forma manual, siguiendo las directrices marcadas por métodos y metodologías de desarrollo de ontologías. Este proceso trata de poner de acuerdo a un conjunto de personas, expertos e ingenieros ontológicos, en la modelización de un dominio concreto, lo que supone un gasto en tiempo y recursos muy alto. Un enfoque interesante que intenta reducir el tiempo y recursos utilizados durante este proceso es la creación de ontologías de forma automática (*Ontology Learning*) en el que el objetivo principal es el de crear ontologías a partir de un corpus de documentos, haciendo uso de técnicas y métodos de campos como el aprendizaje automático, la recuperación de información o el procesamiento del lenguaje natural.

Como en todo proceso de creación de sistemas inteligentes de forma automática, uno de los apartados más importantes que se debe llevar a cabo es la validación o evaluación de dichos sistemas con el fin de comprobar que los modelos representan el conocimiento deseado. Durante la última década no se han realizado avances significativos en la fase de validación del aprendizaje ontologías. A lo largo de este trabajo se mostrará un estudio pormenorizado de las ideas, técnicas y métodos que se han propuesto en el estado de arte y se diseñará e implementará un nuevo método de evaluación que tendrá como objetivo principal la estandarización en el apartado de validación de los procesos de aprendizaje de ontologías. El método propuesto se basa en técnicas cuantitativas con el fin de realizar un proceso lo más automático posible y poder así, de forma rápida y sencilla, comprobar que el conocimiento del dominio representado en el corpus de documentos utilizados para el aprendizaje se refleja en la ontología aprendida de forma automática.

Abstract

Since its invention, ontological creation has been carried out mostly manually, following the guidelines set by the methods and methodology of ontology development. This process involved getting a certain group of people, experts and ontological engineers, to agree into the modelization of a specific domain, which meant a large expenditure of time and resources. An interesting approach that tries to reduce the amount of time and resources used during this rudimentary process is Ontology Learning, in which the main objective is creating ontologies from a corpus of documents, using technics and methods such as machine learning, information retrieval or natural language processing. As any process of automatic creation of intelligent systems, one of the most important steps to be done is the evaluation of these systems in order to ensure that the models represent the desired knowledge. Unfortunately, there is no significant progress in the evaluation of Ontology Learning during the last decade.

Over the course of this paper a detailed study of the ideas, techniques and methods proposed in the state of art will be displayed, and it will also design and implement a new method of evaluation in order to standardize the evaluation of ontology learning processes. The proposed method is based on quantitative techniques that will allow us to perform the process as automatically as possible, so we can quickly and easily verify that the domain knowledge represented in the corpus of documents used for learning is reflected in the automatically learnt ontology.

Índice general

1. Introducción **1**
1.1. Estructura del documento 2

2. Estado del Arte **5**
2.1. Métodos de evaluación cualitativos 6
2.2. Métodos de evaluación cuantitativos 7
2.2.1. Técnicas para la evaluación de la capa léxica 8
2.2.2. Técnicas para la evaluación de la capa taxonómica 10

3. Objetivos **19**
3.1. Objetivo general 19
3.2. Objetivos específicos 19

4. Diseño **21**
4.1. Fase 1: Evaluación de la capa léxica 21
4.1.1. Construcción del *Gold Standard* léxico 22
4.1.2. Método de evaluación para la capa léxica 24
4.2. Fase 2: Construcción de relaciones taxonómicas 25
4.3. Fase 3: Evaluación de la capa taxonómica 26
4.3.1. Evaluación global 27
4.3.2. Evaluación local 27

5. Implementación y experimentación **31**
5.1. El dominio de "*Computer Graphics*" 31
5.2. Uso de herramientas para la generación de *Gold Standards* 32
5.2.1. CrowdFlower 32
5.3. Resultados 34
5.3.1. Evaluación fase 1 35
5.3.2. Nivel de acuerdo entre expertos 36

6. Contribuciones y trabajo futuro **39**
6.1. Contribuciones 39
6.2. Trabajo Futuro 40

A. Artículo EKAW2016 **43**

Bibliografía **65**

Índice de figuras

2.1. Ontologías de ejemplo [Dellschaft and Staab, 2008] 12

4.1. Diseño general de la evaluación 22
4.2. Diseño de la primera fase del proceso de evaluación 23
4.3. Ejemplo del cálculo de Fleiss' Kappa 25
4.4. Diseño de la segunda fase del proceso de evaluación 26
4.5. Diseño de la tercera fase del proceso de evaluación 27
4.6. Ejemplo de jerarquía taxonómica estándar (Or) 28
4.7. Ejemplo de jerarquía aprendida automáticamente (Oc) 29

5.1. Edición del dataset a través de CrowdFlower 33
5.2. Diseño del experimento a través de CrowdFlower 34
5.3. Captura del experimento realizado en CrowdFlower 35

Índice de tablas

2.1. Relación de enfoques propuesta por [Brank et al., 2005] 6
2.2. Semantic cotopies [Dellschaft and Staab, 2008] 13
2.3. Common semantic cotopies [Dellschaft and Staab, 2008] 13

4.1. Medición del *common semantic cotopy* 29

5.1. Medidas de evaluación de la capa léxica 36
5.2. Medidas de precisión para de la capa léxica 36
5.3. Nivel de acuerdo entre expertos 37

Capítulo 1

Introducción

Una ontología se define como: "Una conceptualización compartida formal y explícita sobre un dominio de interés" [Studer et al., 1998]. Este tipo de formalización se planteó como un procedimiento encargado de resolver el cuello de botella que surgía en la adquisición del conocimiento para la construcción de sistemas inteligentes [Studer et al., 1998, Fernández-López et al., 1997].

A pesar de las ventajas del uso de ontologías en lo que respecta a la reducción de tiempo y esfuerzo en la adquisición de conocimiento, su creación sigue siendo un proceso costoso. El procedimiento es largo y, normalmente, se realiza de forma manual entre un conjunto de personas que se deben poner de acuerdo en la modelización de un dominio [Pinto and Martins, 2004], siguiendo o no métodos o metodologías de construcción de ontologías. Para tratar con este problema se planteó la posibilidad de crear las ontologías de forma automática a partir de un conjunto de datos o documentos. Esta idea, que se acuñó con el término *Ontology Learning*, trata de, haciendo uso de algoritmos de aprendizaje automático, construir, a partir de un corpus de datos o documentos, un modelo ontológico que represente el conocimiento descrito en dichas fuentes.

Existen múltiples técnicas o enfoques para implementar procesos de *Ontology Learning*. Por una parte, existen las técnicas basadas en la estadística que se derivan de propuestas realizadas en campos como la recuperación de información o la minería de datos. Se tratan de enfoques muy utilizados en los pasos iniciales del aprendizaje de la ontología para la extracción de términos y jerarquías [Velardi et al., 2005, Turney, 2001]. Por otra parte, las técnicas basadas en lingüística se basan principalmente en procesos de procesamiento del lenguaje natural son, también, muy utilizadas en procesos de *Ontology Learning* [Brill, 1992, Schmid, 2013]. Por último, las técnicas basadas en lógica son menos utilizadas que las anteriores ya que abordan problemas más complejos como la identificación de relaciones o axiomas [Lavrac and Dzeroski, 1994, Shamsfard and Barforoush, 2004].

Como en cualquier proceso de *Machine Learning*, se debe realizar la evaluación

del modelo aprendido con el fin de asegurarse que dicho modelo concuerda con el conocimiento que se desea representar. Este proceso de evaluación es la base de esta tesis de fin de máster, en la que se intentarán abordar las cuestiones más relevantes del campo desde un punto de vista científico.

El proceso de evaluación de ontologías aprendidas automáticamente fue estudiado con bastante profundidad a mediados de la década de los 2000 [Maedche and Staab, 2002, Brank et al., 2005, Porzel and Malaka, 2004, Sabou et al., 2005], sin embargo, en los últimos años no se han realizado avances significativos. La principal razón por la que no se ha progresado en este campo, es que, de la misma manera que en el aprendizaje de los modelos, estos procesos combinan múltiples ideas y técnicas que provienen de diversos campos de investigación (recuperación de la información, procesamiento del lenguaje natural, etc.) y que durante esa época no habían sido desarrollados lo suficiente como para poder ofrecer unas valoraciones adecuadas sobre la relación del modelo aprendido y el conocimiento del dominio que se deseaba representar.

Actualmente, los avances en estos campos se han utilizado para mejorar y desarrollar métodos en el apartado del aprendizaje de los modelos, sin tener muy en cuenta la parte de la evaluación y, es por eso, que se ha detectado la necesidad de proponer un nuevo método de evaluación que complete el proceso de *Ontology Learning*. Por esta razón la propuesta desea ocupar el vacío existente en la evaluación de *Ontology Learning*, ofreciendo un sistema lo más independiente y automático posible. Además, en este caso en particular, la propuesta que se realizará se centrará en la evaluación de ontologías de dominio, no teniendo en cuenta otros tipos de modelos como las *Upper Ontologies* o *Lightweight Ontologies*.

Por último, cabe destacar que esta tesis de fin de máster se encuentra en el contexto de un proyecto europeo *DrInvetor: Your personal research assistant*[1]. El objetivo principal de este proyecto es el de implementar un sistema de recomendación de artículos científicos en el dominio de los gráficos por computador haciendo uso de modelos de representación del conocimiento como son las ontologías. Debido a que se trata de un dominio muy extenso y complejo y no existen ontologías robustas y testeadas se ha optado por utilizar técnicas de *Ontology Learning* para construir estos modelos.

1.1. Estructura del documento

El trabajo que se presenta a lo largo del documento se distribuye de la siguiente forma:

- En el **capítulo dos** se describirá el estado del arte de la evaluación de on-

[1] `http://drinventor.eu/`

tologías aprendidas poniendo especial atención a los métodos de evaluación cuantitativos, entendidos como métodos que ofrecen valores numéricos calculados a partir de diferentes fórmulas sobre la evaluación, por lo que están más enfocados a la automatización.

- En el **capítulo tres** se detallará el objetivo general de esta tesis fin de máster y, de forma más pormenorizada, los objetivos específicos.
- En el **capítulo cuatro** se presentará un nuevo método de evaluación de ontologías basado en la creación de *Gold standards* a partir de expertos humanos y técnicas de evaluación cuantitativas.
- En el **capítulo cinco** se describirá el experimento llevado a cabo para evaluar ontologías haciendo uso del método propuesto en el capítulo anterior. La experimentación se desarrollará a partir de un corpus de artículos científicos en el dominio de los gráficos por computador.
- En el **capítulo seis** se describirán las contribuciones realizadas por este trabajo, una serie de conclusiones y se detallarán algunas de las posibles líneas futuras de investigación.

Capítulo 2

Estado del Arte

En el estado del arte, se han planteado muchos tipos de clasificaciones de los diferentes métodos y técnicas de evaluación del aprendizaje de ontologías. Por ejemplo, en [Wong et al., 2012] se realiza una clasificación bastante simple, obviando muchos métodos y técnicas relevantes, en función de la estrategia de evaluación a llevar a cabo: basado en tareas, basado en corpus o basada en criterios. En [Dellschaft and Staab, 2008] se propone otro tipo de clasificación que engloba a la anteriormente citada, basada en el escenario de actuación de las estrategias de evaluación: comparando algoritmos de aprendizaje o estimando la calidad de la ontología observando si cumple el objetivo para la que fue creada.

Por lo general, al realizar la clasificación, surgen dos planteamientos bastante diferenciados. Por una parte, el enfoque centrado en qué capa de la ontología se debe evaluar: a) la capa léxica o conceptual, en la que se evalúan los términos y conceptos que forman el modelo; la capa taxonómica, en la que se observa si las relaciones de clasificación aprendidas son las correctas; c) o la capa no-taxonómica, en la que se evalúan este otro tipo de relaciones. Por otra parte, esta el planteamiento que se centra en dividir las técnicas en función de la estrategia que se desea llevar a cabo (similar a lo planteado en [Wong et al., 2012] pero incluyendo otras estrategias). Los dos planteamientos se complementan uno con otro debido a que al escoger una estrategia se pueden evaluar varias capas de una ontología y viceversa, para evaluar una capa se pueden utilizar varias estrategias y combinarlas. En [Brank et al., 2005] se presentan las relaciones entre los dos planteamientos, como se puede observar en la Figura 2.1.

En este documento se presenta una clasificación que pretende englobar todos los métodos y técnicas existentes desde un punto de vista que no se ha abordado de momento. El problema de los planteamientos anteriores es que dejan en un segundo plano un punto que se debería tener muy en cuenta: la automatización del proceso.

La mayoría de los métodos propuestos carecen de automatización, y son ne-

	Approach to evaluation			
Level	Gold standard	Application-based	Data-driven	Assessment by humans
Lexical, vocabulary, concept, data	x	x	x	x
Hierarchy, taxonomy	x	x	x	x
Other semantic relations	x	x	x	x

Tabla 2.1: Relación de enfoques propuesta por [Brank et al., 2005]

cesarias personas expertas para poder llevar a cabo el proceso. Es por eso que se describe una clasificación que dividirá los métodos entre los que son de tipo cualitativo, por lo tanto, muy complicados de automatizar, y los que son de tipo cuantitativo.

Debido a que el método que se especifica en el Capítulo 4 es un método que trata de dotar al proceso de la mayor automatización posible, el estado del arte descrito a continuación se centrará en mayor medida en técnicas que permitan realizar la evaluación de manera que no sea necesaria la intervención de expertos. En el apartado de evaluación cuantitativa se detallarán las técnicas y estrategias que permitan evaluar tanto la capa léxica como la capa taxonómica, no teniendo en cuenta la evaluación de la capa no-taxonómica, debido a que es un campo muy complejo que no ha sido abordado con la suficiente profundidad en el estado del arte como para que se incluya en dicho documento.

2.1. Métodos de evaluación cualitativos

Las evaluaciones cualitativas del proceso de *Ontology Learning* permiten evaluar la calidad de los modelos aprendidos haciendo uso de los conocimientos de expertos en el dominio que se desea representar. Como se ha comentado anteriormente, estos métodos ralentizan mucho el proceso de evaluación ya que, generalmente, se hacen de forma manual. Aun así, este aspecto también puede ser una ventaja ya que, de esta forma, los resultados no dependen de la calidad de las tecnologías, métodos o técnicas que permitan automatizar la evaluación.

El enfoque más relevante que se ha propuesto ha sido la "Evaluación manual por expertos humanos". Esta estrategia ha sido propuesta como sistema de evaluación en numerosos artículos donde son expertos humanos los que se encargan de juzgar la calidad del modelo construido [Berland and Charniak, 1999, Girju et al., 2002].

El problema de esta propuesta es que tiene muchos inconvenientes con los que es difícil tratar:

1. El modelo obtenido se compara con el conocimiento de un experto, no con la información que se encuentra en el corpus.

2. Determinar en qué medida el conocimiento identificado en el modelo coincide con el conjunto del conocimiento que se debería haber identificado a partir del corpus. En recuperación de la información, esta medida se conoce como *recall*.

3. La evaluación dependerá de qué experto o expertos la realicen. Al existir diferentes puntos de vista sobre la representación del conocimiento en un dominio esto conduciría a diferentes evaluaciones y resultados.

4. El coste de usar expertos para realizar la evaluación es muy alto.

Al analizar los puntos anteriores en profundidad, se puede inferir que realizar un proceso de evaluación de forma cualitativa es costoso y aunque, pueden generar buenos resultados si se diseña y desarrolla de forma correcta, un enfoque cuantitativo permitiría disminuir los costes de evaluación, generar un proceso con capacidad para ser automatizado y obtener unos resultados similares a un proceso de evaluación que involucrase expertos humanos. Algunos ejemplos de evaluaciones cualitativas se encuentran descritas en [Gómez-Pérez, 1999, Guarino, 1998].

2.2. Métodos de evaluación cuantitativos

En los métodos de evaluación cuantitativa se propone ofrecer un feedback numérico y objetivo en función de la calidad de la ontología aprendida. Este tipo de métodos, además de estar más enfocados a la automatización, permiten obtener una evaluación que resuelve algunos de los problemas que se han reflejado anteriormente en el caso de hacer uso de métodos cualitativos. Los enfoques o estrategias más relevantes son:

- **Evaluación a partir de un *Gold Standard*:** Se trata de crear un modelo que sirva para evaluar el modelo aprendido realizando una comparación con un modelo base o estándar. Cuanto más se parezca el modelo aprendido al *Gold Standard*, se considerará que su calidad será mayor. Tanto para la capa léxica como la taxonómica se utilizan diferentes técnicas que se detallarán a continuación. El inconveniente de este método es el proceso de creación del *Gold Standard* que, generalmente, se realizaba de forma manual. Ejemplos de este tipo de método se pueden encontrar en [Dellschaft and Staab, 2008, Cimiano et al., 2004, Spyns and Reinberger, 2005, Sabou et al., 2005].

- **Evaluación basada en tareas:** Se trata de evaluar un modelo a partir de su comportamiento al ser aplicado en la realización de una o varias tareas. Normalmente se trata de tareas como recuperación de información o búsqueda de conceptos entre otras, y se requiere la creación de un *Gold Standard* de respuestas, lo que dificulta la posibilidad de automatización. La propuesta más relevante para este enfoque fue planteada por [Porzel and Malaka, 2004].

- **Evaluación basada en corpus:** Este tipo de enfoque se utiliza para comprobar la cobertura que ofrece un modelo de un dominio. Para ello se realizan operaciones de extracción de información en un corpus de documentos y se compara con el modelo aprendido. Uno de los mayores inconvenientes de este procedimiento es que hace uso de métodos de *clustering* y procesamiento del lenguaje natural, lo que hace que los resultados obtenidos dependan en gran medida de la calidad de las tecnologías utilizadas para realizar la extracción de información. Ejemplos de aplicación de este tipo de evaluación sobre la capa léxica se pueden encontrar en [Daelemans and Reinberger, 2004, Brewster et al., 2004].

- **Evaluación basada en criterios:** En este tipo de enfoque se trata de obtener medidas de evaluación que permitan calcular de qué manera un modelo ontológico puede asociarse a una serie de criterios que pueden ser tanto relacionados con la estructura de la ontología [Gangemi et al., 2006, Gómez-Pérez, 2004] o evaluaciones algo más sofisticadas basadas en, por ejemplo nociones filosóficas [Guarino and Welty, 2009]. Este tipo de métodos son muy complicados de automatizar ya que necesitan la aportación de ingenieros ontológicos y expertos del dominio para poder llevarse a cabo, por lo no se describirán en los apartados siguientes.

En los apartados que se detallan a continuación se presentan las técnicas más relevantes en la disciplina de las evaluaciones cuantitativas clasificadas en función de la capa del modelo que se desea evaluar: capa léxica o capa taxonómica.

2.2.1. Técnicas para la evaluación de la capa léxica

Las técnicas de evaluación para la capa léxica permiten determinar si los conceptos aprendidos durante el proceso de *Ontology Learning* son correctos y reflejan el conocimiento del corpus de documentos que ha sido usado durante el aprendizaje. Para ello, y, asociadas a los métodos detallados en el apartado anterior, se han propuesto diferentes técnicas que se describen a continuación:

2.2.1.1. Técnicas basadas en *Gold Standard*

Generalmente estás técnicas se basan en medidas que provienen del ámbito de la recuperación de información. La idea de adaptar estas medidas para la evaluación de ontologías se presenta en [Sabou et al., 2005] y posteriormente se proponen una serie de mejoras en [Dellschaft and Staab, 2008]. Las medidas que se detallan a continuación se obtienen a partir de la comparación del modelo ontológico obtenido con un modelo de referencia:

- **Recall**: Determina en qué medida el conocimiento identificado en el modelo coincide con el conjunto del conocimiento que se debería haber identificado a partir del corpus. Como las fórmulas propuestas en [Sabou et al., 2005] y en [Dellschaft and Staab, 2008] son muy similares, en la ecuación 2.1 se muestra la fórmula planteada en este último, donde Comp se refiere a los conceptos aprendidos y Ref a los conceptos de referencia.

$$Recall(Ref, Comp) = \frac{|Comp \bigcap Ref|}{|Ref|} \tag{2.1}$$

- **Precisión**: Determina en qué medida el conocimiento reflejado en la ontología se ha identificado correctamente, es decir, de todos los conceptos que forman parte del modelo, cuales tienen relación con el conocimiento que se quiere representar. De la misma forma que en el apartado de Recall, en la ecuación 2.2 se muestra la formulada descrita en [Dellschaft and Staab, 2008].

$$Precision(Ref, Comp) = \frac{|Comp \bigcap Ref|}{|Comp|} \tag{2.2}$$

- **F-Measure:** Esta medida relaciona los resultados obtenidos en recall y precisión. La fórmula general se muestra en la ecuación 2.3 pero en la mayoría de propuestas, como en [Dellschaft and Staab, 2008], el valor de β suele ser 1.

$$F_{\beta-score} = \frac{(1+\beta^2)(precision \times recall)}{\beta(precision + recall)} \tag{2.3}$$

2.2.1.2. Técnicas basadas en tareas

Las técnicas propuestas con este enfoque no se han desarrollado tanto como las técnicas basadas en *Gold Standard*. La propuesta más relevante se realiza en [Porzel and Malaka, 2004], en la cual a partir de la aplicación del modelo aprendido a una tarea específica se analizan las respuestas ofrecidas por dicho modelo y

se comparan con un estándar de respuestas (realizadas por un humano). Para cuantificar los errores que comete la ontología aprendida se estudia cómo afectan las inserciones, eliminaciones y sustituciones de conceptos a las respuestas que ofrecen. Aunque se trata de un enfoque prometedor, además de que no existe en la literatura ninguna descripción de técnicas específica que permita llevar a cabo la idea de evaluación que se propone, este enfoque requiere que el proceso se centre de forma muy específica en el dominio a evaluar. Esta restricción genera un coste muy alto en el curso de la evaluación ya que se deberían tener en cuenta a lo largo del proceso un conjunto de expertos capaces de realizar evaluaciones muy concretas y específicas. Y además, al crear ontologías de forma automática, la identificación de un dominio de forma tan precisa a partir del corpus de documentos es una tarea muy compleja.

2.2.1.3. Técnicas basadas en corpus

La propuesta más relevante que hace uso de un corpus de documentos para realizar la evaluación se presenta en [Brewster et al., 2004]. En ella se describe un enfoque para una evaluación probabilística basada en el teorema de Bayes. A partir de un corpus anotado buscan entre un conjunto de ontologías cual de ellas obtiene la mayor probabilidad a posteriori dado el corpus (ecuación 2.4). La ventaja de este planteamiento es que crea una medida objetiva de la calidad del modelo. Aún así, esto, como se ha comentado anteriormente, es también una desventaja ya que depende directamente del comportamiento de las técnicas de extracción de información utilizadas sobre el corpus que sirven como una especie de *Gold Standard* para realizar la evaluación.

$$O^* = argmax_o P(O|C) = argmax_o \frac{P(C|O)P(O)}{P(C)} \tag{2.4}$$

Como se ha podido observar en los apartados anteriores, el concepto de *Gold Standard* para la evaluación de la capa léxica es un término constante en todos los enfoques propuestos, por lo que es un planteamiento a tener en cuenta en la implementación de nuevas propuestas para la evaluación de *Ontology Learning*.

2.2.2. Técnicas para la evaluación de la capa taxonómica

Las técnicas de evaluación que permiten determinar la calidad de las relaciones taxonómicas en un modelo ontológico han utilizado, de la misma manera que en el apartado anterior, diferentes planteamientos o enfoques para desarrollar la

evaluación. Generalmente, para evaluar esta capa se realizan dos tipos de medidas: por una parte, en la evaluación local, se compara las similitudes entre la posición de dos conceptos entre una jerarquía aprendida y otra de referencia, mientras que en la evaluación global se calcula el promedio de las medidas locales para poder ofrecer un único resultado para toda una ontología. Aunque se trata de una tarea compleja y en la que no se ha conseguido avanzar demasiado, a continuación se detallan las propuestas más relevantes en el estado del arte para la realización de este proceso.

2.2.2.1. Técnicas basadas en *Gold Standard*

Estas propuestas que hacen uso de este tipo de enfoque han sido las más desarrolladas en el estado del arte. Existen planteamientos bastante antiguos como el detallado en [Hahn and Schnattinger, 1998], que se actualizó posteriormente en [Maynard et al., 2006, Maynard et al., 2008]. Aunque se trata de enfoques interesantes, este apartado se centrará en la propuesta realizada en [Dellschaft and Staab, 2008] en la que se definió un framework para medir la calidad de las relaciones taxonómicas bastante completo y en la propuesta de [Spyns, 2005] en la que se evalúan tripletas y se construye el *Gold Standard* de forma automática.

Framework taxonómico

Para poder comprender la propuesta de [Dellschaft and Staab, 2008] lo primero que se debe definir es lo que los autores han descrito como el *core ontology*. Una vez definido dicho concepto se describirán las diferentes partes que conforman el framework para la medición de la calidad de las relaciones taxonómicas.

Lema: *La estructura* $O := (C, root, \leq c)$ *se denomina como el concepto core ontology.* C *es el conjunto de conceptos identificados y root es definido como el concepto raíz que cumple el orden parcial* $\leq c$ *en* C*. Este orden parcial es llamado jerarquía de conceptos o taxonomía. La ecuación* $\forall c \in C : c \leq croot$*, se cumple para las jerarquías de conceptos a evaluar.*

Precisión y recall taxonómicos: A lo largo de esta subsección se describirán las medidas de precisión y recall para relaciones taxonómicas. Los autores únicamente han definido la medida de precisión ya que el recall se puede derivar de la anterior de forma muy fácil. Además, este apartado mejora y extiende el framework desarrollado en [Maedche and Staab, 2002] donde se definía la idea sobre la medida *taxonomic overlap* que se definirá en apartados siguientes.

- *Comparación de conceptos*
  Como se ha mencionado anteriormente, estos procesos de medición se dividen en dos apartados: las medidas de carácter local y las medidas globales. En

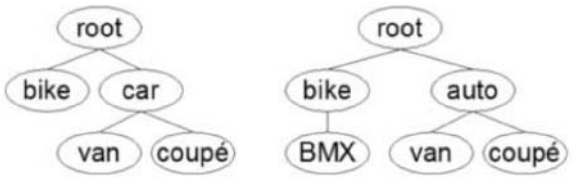

Figura 2.1: Ontologías de ejemplo [Dellschaft and Staab, 2008]

este caso, las medidas de carácter local se encargan de comparar la posición de dos conceptos en una jerarquía, mientras que la medida global es usada para comparar dos jerarquías completas.

Para medir la relación entre dos conceptos en una jerarquía de forma cuantitativa se puede realizar el cálculo propuesto en la ecuación 2.5. En esta ecuación se calcula la precisión taxonómica local (tp_{ce}) en función de una característica propia de los conceptos definidos (ce) en la jerarquía, siendo $c_1 \in O_C$ y $c_2 \in O_R$ los conceptos utilizados para calcular la medida.

$$tp_{ce}(c_1, c_2, O_C, O_R) := \frac{|ce(c_1, O_c) \cap ce(c_2, O_R)|}{|ce(c_1, O_c)|} \tag{2.5}$$

Una de las características (ce) que se han propuesto a lo largo del estado del arte para calcular esta medida fue el *semantic cotopy* ([Maedche and Staab, 2002]). Esta idea sugiere caracterizar los conceptos a partir de todos sus superconceptos y subconceptos. El *semantic cotopy* se define en la ecuación 2.6.

$$sc(c, O) := \{c_i | c_i \in C \ \wedge \ (c_i \leq c \vee c \leq c_i)\} \tag{2.6}$$

Debido a que las medidas de precisión y recall no son independientes entre sí, y estas medidas de evaluación deben ser estimadas a partir de medidas no dependientes, para calcular la precisión taxonómica local no se recomienda utilizar como característica asociada a los conceptos el *semantic cotopy*. Este problema de dependencias se puede resolver haciendo uso de una extensión de la idea anterior. El *common semantic cotopy*(csc), definido en la ecuación 2.7, no tiene en cuenta en su cálculo los conceptos que no están disponibles en el otro modelo ontológico (C_2) por lo que se resuelve el inconveniente del *semantic cotopy*.

$$csc(c, O_1, O_2) := \{c_i | c_i \in C_1 \cap C_2 \ \wedge \ (c_i <_1 c \vee c <_1 c_i)\} \tag{2.7}$$

En las tablas 2.2 y 2.3 se pueden observar las diferencias entre las dos ideas descritas anteriormente con un ejemplo que hace referencia a los cálculos que se deberían realizar en el caso de que se quisiesen obtener las medidas en referencia de las ontologías definidas en la Figura 2.1.

c	$sc(c, \mathcal{O}_{R1})$	$sc(c, \mathcal{O}_{C1})$
root	{root, bike, car, van, coupé}	{root, bike, BMX, auto, van, coupé}
car	{root, car, van, coupé}	–
auto	–	{root, auto, van, coupé}
van	{root, car, van}	{root, auto, van}
coupé	{root, car, coupé}	{root, auto, coupé}
bike	{root, bike}	{root, bike, BMX}
BMX	–	{root, bike, BMX}

Tabla 2.2: Semantic cotopies [Dellschaft and Staab, 2008]

c	$csc(c, \mathcal{O}_{R1}, \mathcal{O}_{C1})$	$csc(c, \mathcal{O}_{C1}, \mathcal{O}_{R1})$
root	{bike, van, coupé}	{bike, van, coupé}
car	{root, van, coupé}	–
auto	–	{root, van, coupé}
van	{root}	{root}
coupé	{root}	{root}
bike	{root}	{root}
BMX	–	{root, bike}

Tabla 2.3: Common semantic cotopies [Dellschaft and Staab, 2008]

- *Comparación de jerarquías de conceptos*

  Una vez detalladas las medidas locales, se puede definir el marco de referencia para la construcción de la medida de precisión taxonómica global. En la ecuación 2.8 se define la forma de obtener dicha medida, que está influenciada por la precisión obtenida en la capa léxica. Para cada concepto de la jerarquía se calcula la precisión taxonómica local y posteriormente se calcula la media. Existen dos maneras diferentes de calcular la precisión taxonómica. Por una parte, la primera opción es la detallada en el apartado anterior y se ejecutará en el caso en el que el concepto del cual se quiere obtener la precisión taxonómica se encuentre en la ontología de referencia. La segunda opción se ejecutará en el caso de que el concepto a evaluar no se encuentre en el *Gold Standard*, en cuyo caso se debe realizar una estimación sobre su precisión taxonómica. Para realizar esta estimación, en [Maedche and Staab, 2002] se propone comparar el concepto de la ontología aprendida con todos los conceptos de referencia y hacer uso del valor más alto de la precisión taxonómica local.

  Todos estos cálculos, como se ha comentado anteriormente, están influencia-

dos por la precisión obtenida en el proceso de evaluación de la capa léxica. Para eliminar dicha influencia, únicamente se deberían tener en cuenta aquellos conceptos que aparecen descritos en las dos ontologías. En ese caso se utilizarían, tanto para la precisión como para el recall, las ecuaciones 2.9 y 2.10 que hacen uso del *common semantic cotopy*.

$$TP(O_C, O_R) := \frac{1}{|C_C|} \sum_{c \in C_C} \begin{cases} tp(c, c, O_C, O_R) & \text{if } c \in C_r \\ max_{c' \notin C_R}\, tp(c, c', O_C, O_R) & \text{if } c \notin C_R \end{cases} \tag{2.8}$$

$$TP_{csc}(O_C, O_R) := \frac{1}{|C_C \cap C_R|} \sum_{c \in C_C \cap C_R} tp_{csc}(c, c, O_C, O_R) \tag{2.9}$$

$$TR_{csc}(O_C, O_R) := TP_{csc}(O_R, O_C) \tag{2.10}$$

F- y F-Measure Taxonómico: Como en el caso de la capa léxica, en la capa taxonómica también existe una medida que relaciona la precisión y el recall. La *taxonomic F-measure* se detalla en la ecuación 2.11 y se trata de la media armónica entre la precisión taxonómica global y el recall.

$$TF(O_C, O_R) := \frac{2 \cdot TP(O_C, O_R) \cdot TR(O_C, O_R)}{TP(O_C, O_R) + TR(O_C, O_R)} \tag{2.11}$$

Si se ha decidido no tener en cuenta la influencia de la evaluación de la capa léxica en pasos anteriores del proceso, sería recomendable hacer uso de la ecuación 2.12 para obtener la media armónica entre la TF y la precisión léxica, con el fin de obtener unos resultados que reflejen la calidad de la ontología de forma objetiva.

$$TF'(O_C, O_R) := \frac{2 \cdot LR(O_C, O_R) \cdot TF(O_C, O_R)}{LR(O_C, O_R) + TF(O_C, O_R)} \tag{2.12}$$

Overlap Taxonómico: En [Cimiano et al., 2004, Maedche, 2012] se describe una nueva medida que intenta tratar con el problema de la asimetría entre las ontologías. La parte global se calcula de la misma manera que en TP pero la parte local se calcula como se describe en la ecuación 2.13.

$$to_{sc}(c_1, c_2, O_1, O_2) = \frac{|sc(c_1, O_1) \cap sc(c_2, O_2)|}{|sc(c_1, O_1) \cup sc(c_2, O_2)|} \tag{2.13}$$

Los autores definen esta idea como una especie de precisión and recall sobre ontologías asimétricas, pero se ha observado en [Dellschaft and Staab, 2008] que las medidas detalladas con anterioridad son las que ofrecen una evaluación de calidad y las que deben ser usadas para el proceso en la capa taxonómica.

Evaluación de tripletas

La idea principal de esta propuesta es la de evaluar de forma léxica tripletas generadas automáticamente a partir del corpus de documentos. Para ello presenta una serie de fórmulas que permiten medir el rendimiento, la precisión y el *recall*, así como la cobertura de un dominio. Todo ello a partir de la comparación de estas tripletas con un *Gold Standard* que se genera de forma automática. El inconveniente de esta propuesta es dicha generación automática del modelo de referencia, ya que, al hacer uso de ecuaciones estadísticas muy simples producen muchos errores en la extracción del conocimiento. Es por eso que los resultados descritos en [Spyns, 2005] no se deben tener en cuenta ya que no queda muy claro si son debido a una mala creación del *Gold Standard* o a la calidad de las técnicas propuestas.

$$\tilde{f} = (\frac{f_{word\_text}}{N}) * 100 \tag{2.14}$$

$$z = \frac{\tilde{f}_1 - \tilde{f}_2}{\sqrt{(\frac{\tilde{f}_1 * (100 - \tilde{f}_1)}{N_1}) + (\frac{\tilde{f}_2 * (100 - \tilde{f}_2)}{N_2})}} \tag{2.15}$$

Para generar automáticamente el *Gold Standard* se utilizan las ecuaciones 2.14 y 2.15. En la ecuación 2.14 la variable f_{word_text} hace referencia a la frecuencia absoluta de una palabra en un texto y N al número total de palabras del texto. Por otra parte, en la ecuación 2.15 se calcula el conjunto de palabras relevantes en el texto de forma estadística comparando las desviaciones de los valores calculados en la ecuación 2.14 entre las diferentes palabras.

$$coverage(triples, text) = \frac{\sum_{i=1}^{n} \frac{\#(words\_triples\_freq\_class_i \cap words\_text\_freq\_class_i)}{words\_text\_freq\_class_i} * 100}{n} \tag{2.16}$$

Para calcular la cobertura que realizan las tripletas aprendidas sobre el dominio se hace uso de la ecuación 2.16. En esta ecuación las palabras del texto son agrupadas en *frequency classes*, esto es, la cantidad total de veces que una palabra aparece en el corpus. Para cada una de estas *frequency classes* se calcula su intersección con las palabras de las tripletas y finalmente, se realiza una media

sobre el número total de clases.

$$recall(triples, text) = \left(\frac{\#(words\_of\_triples\_mined \cap statistically\_relevant\_words)}{\#statistically\_relevant\_words}\right) * 100 \quad (2.17)$$

$$precision(triples, text) = \frac{\#(words\_of\_triples\_mined \cap statistically\_relevant\_words)}{\#words\_of\_triples\_mined}) * 100 \quad (2.18)$$

La idea de precisión y recall propuesta es muy similar a la detallada en la capa léxica, salvo que se hace uso de la palabras que forman las tripletas.

$$accuracy(triples, text) = \frac{\sum_{i=1}^{n} \frac{\#(words\_triples\_rel\_freq\_class_i \cap words\_text\_rel\_freq\_class_i)}{words\_text\_rel\_freq\_class_i} * 100}{n} \quad (2.19)$$

La medida del rendimiento, reflejada en la ecuación 2.19, es muy similar a la presentada en la ecuación 2.16, pero en vez de hacer uso de todas las *frequency classes*, únicamente tiene en cuenta las clases más relevantes, en este caso, se consideran las *frequency classes* que contengan más de un 60 % del vocabulario destacado.

2.2.2.2. Técnicas basadas en tareas

Aunque no existen técnicas asociadas a este enfoque en el estado del arte, la idea desarrollada por [Porzel and Malaka, 2004] podría servir tanto para evaluar la capa léxica como para evaluar la capa taxonómica, ya que se trata de analizar cómo se comporta un modelo ontológico en la realización de una tarea.

2.2.2.3. Técnicas basadas en corpus

Aunque la idea que propone [Spyns, 2005] podría relacionarse con este enfoque, ya que hace uso de tecnologías de procesamiento del lenguaje natural para realizar extracción del conocimiento, el fin de ese proceso es el de crear un *Gold Standard* de manera automática, por lo que se ha optado por incluirla en ese apartado.

Como se puede observar, estás dos últimas técnicas no permiten realizar una evaluación en la capa taxonómica de forma exhaustiva. Se puede asegurar que hoy en día, si se desea realizar una evaluación correctamente, el uso de un *Gold Standard* para evaluar relaciones taxonómicas es casi obligado. Como se ha visto en los diferentes enfoques detallados en esta sección, todas las propuestas hacen uso de estos modelos estándar, ya sean en forma de ontología [Dellschaft and Staab, 2008] o en forma de respuestas a una tarea [Porzel and Malaka, 2004].

Si se analiza la evaluación que realizan las herramientas más relevantes en el campo de *Ontology Learning* se puede observar que todas hacen uso de un Gold Standard. Por ejemplo, en la herramienta Text2Onto [Maedche and Volz, 2001], crean un *Gold Standard* de forma manual haciendo uso de expertos a partir del corpus de documentos utilizado en el proceso de aprendizaje. Otra forma de abordar la evaluación es la que se propone en [Manzano-Macho et al., 2008], en la que se utiliza una ontología contrastada como Gold Standard.

Capítulo 3

Objetivos

A lo largo de este capítulo se describirá, por una parte, el objetivo principal de la tesis fin de máster en el campo de la evaluación de ontologías aprendidas de forma automática, y por otra, una serie de objetivos específicos en cada uno de los apartados a desarrollar a lo largo de este trabajo.

3.1. Objetivo general

El objetivo principal de la tesis fin de máster es la de crear un sistema que realice la evaluación de ontologías que se han construido haciendo uso de algoritmos de aprendizaje automático y técnicas de procesamiento de lenguaje natural. Como se ha comentado en apartados anteriores, la idea de la evaluación se ha pensando con el fin de comprobar de qué forma el modelo ontológico representa el conocimiento del dominio que se desea reflejar a partir del corpus de documentos usado para el aprendizaje. Además, el proceso debe tener como base técnicas de evaluación cuantitativas para la capa léxica y taxonómica y ser lo más automático posible.

3.2. Objetivos específicos

Dentro de los objetivos específicos de la tesis fin de máster, se pueden identificar tres secciones que se deben desarrollar:

- **Estado del arte:** Desarrollar un estado del arte de la evaluación de ontologías aprendidas de forma automática actualizado y completo. Incluir, en dicho análisis, técnicas y métodos de evaluación para la capa léxica y la capa taxonómica, tanto cualitativas como cuantitativas. De esta forma se realizará una clasificación novedosa centrada en las posibilidades de automatización del proceso.

- **Diseño del proceso:** Diseñar un proceso innovador de evaluación de ontologías aprendidas que pueda ser utilizado en los próximos años. Se identifican los siguientes apartados:

  1. **Creación de *Gold Standard***: En el método de evaluación propuesto s harán uso de técnicas de evaluación basadas en *Gold Standard*, por lo que, tanto para la evaluación de la capa léxica como para la de la capa taxonómica, se deben crear dos ontologías que sirvan de estándar a partir de los términos extraídos del proceso de aprendizaje. Debido a que la creación de estos modelos necesita la ayuda de expertos en el dominio a evaluar, se deberá incluir un sistema de evaluación sobre el nivel de acuerdo entre los evaluadores.
  2. **Evaluación de la capa léxica**: Tomando como punto de partida las ideas detalladas en [Dellschaft and Staab, 2008] sobre técnicas de evaluación con *Gold Standard* para la capa léxica diseñar un método de evaluación para dicha capa. Las medidas más relevantes a tener en cuenta a la hora de diseñar la estrategia de evaluación serán la precisión y el recall.
  3. **Evaluación de la capa taxonómica**: Del mismo modo que en la capa léxica, el punto de partida serán las ideas descritas en [Dellschaft and Staab, 2008], de las cuales se prestará especial atención a la especificación del *framework* taxonómico. Por lo tanto, las medidas a tener en cuenta para esta capa serán las de precisión y recall taxonómico además de la F-Measure.

- **Implementación y experimentación:** Se tratará de poner en marcha sobre un caso real, centrado en ontologías del dominio de los gráficos por computador, las fases propuestas a lo largo del diseño del proceso. Para ello se hará uso de herramientas como *Crowdflower*[1], se construirán *datasets* en formato CSV, y se analizarán los resultados con la ayuda de una hoja de cálculo que incluya la evaluación del nivel de acuerdo de los expertos.

[1] `http://www.crowdflower.com/`

Capítulo 4

Diseño

En sección se presenta un nuevo marco de trabajo (*framework*) para la evaluación de ontologías aprendidas automáticamente mediante técnicas de *Ontology Learnig*. Esta evaluación se compondrá de tres fases: en la primera, se evaluará la capa léxica a través de la construcción de un *Gold Standard* a partir de expertos humanos y utilizando técnicas de *crowdsourcing*[1]; la segunda fase hará hincapié en la creación de un *Gold Standard* de relaciones taxonómicas, haciendo uso también de expertos humanos y *crowdsourcing*; y por último, la tercera fase se centrará en la evaluación de la capa taxonómica a partir de las ideas propuestas en [Dellschaft and Staab, 2008].

Las ontologías que se van a utilizar a lo largo de la evaluación en este marco de trabajo pertenecen al campo de *Computer Graphics* y se hará uso de expertos en esta disciplina para la generación de los *Gold Standards* de las dos primeras fases. A simple vista parece un método poco automatizado, pero como se ve en la Figura 4.1 cada proceso toma la salida del anterior y hace uso de ella en su tarea. De esta forma, se realiza conjuntamente la evaluación de todas las capas.

La decisión de realizar la evaluación de esta manera es debido a que, como se puede observar en [Spyns, 2005], la creación automática del *Gold Standard* conduce a una evaluación de mala calidad debido a los problemas que hoy en día siguen teniendo las tecnologías de las que se hace uso para realizar estas tareas. En las siguientes secciones describirán de forma detallada las diferentes fases de las que se compone el método.

4.1. Fase 1: Evaluación de la capa léxica

En la primera fase de la evaluación se evaluarán los términos obtenidos durante el proceso de aprendizaje. Para ello, primero se deberá crear un *Gold Standard*

[1] `https://es.wikipedia.org/wiki/Crowdsourcing`

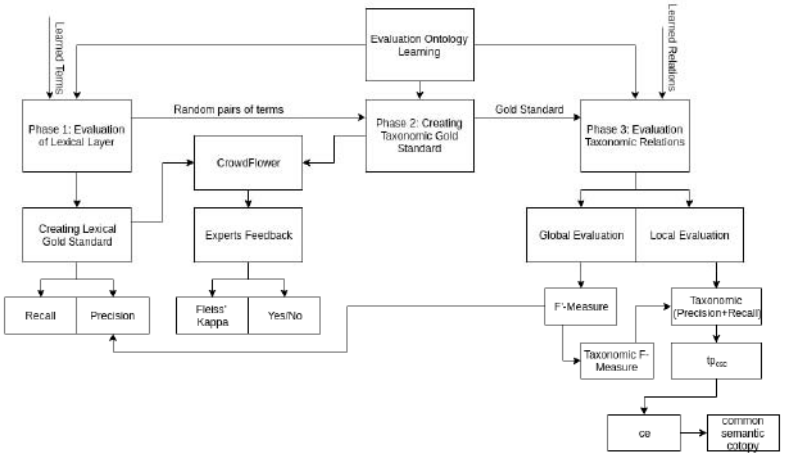

Figura 4.1: Diseño general de la evaluación

de términos y a continuación aplicar técnicas propuestas en el estado del arte [Dellschaft and Staab, 2008, Sabou et al., 2005] para evaluar esta capa. El diseño de esta primera fase se detalla de forma gráfica en la Figura 4.2. A continuación se explicará el método que se ha diseñado para obtener el *Gold Standard* y que técnicas serán las utilizadas para realizar la evaluación.

4.1.1. Construcción del *Gold Standard* léxico

Como se ha observado y comentado en apartados anteriores, el proceso de creación de un *Gold Standard* de forma automática conduce a una evaluación incorrecta debido a los errores que cometen las tecnologías de procesamiento de lenguaje natural actuales usadas en el proceso [Spyns, 2005]. Es por ello que se ha decidido construir un *Gold Standard* a partir de expertos humanos y métodos de *crowdsourcing*. Para ello se hará uso de la herramienta web para análisis de *datasets, CrowdFlower*, que permitirá, a los expertos del domino, a partir de los términos que se han aprendido, analizar si estos pertenecen realmente o no al dominio de *Computer Graphics*. Los expertos no sabrán en ningún momento que los términos que están evaluando han sido aprendidos automáticamente, de esta forma se asegura la objetividad del experimento.

El proceso que deberán realizar los expertos será el de evaluar una serie de palabras contestando simplemente si pertenecen o no al dominio. Por otra parte, antes de comenzar el proceso de evaluación, se le realizarán varias preguntas básicas a cada usuario para comprobar sus conocimientos sobre representación del conocimiento, lenguaje, etc. En el caso de que el usuario no conteste correctamente a estas preguntas, los datos aportados por el mismo serán descartados. Por último, cada término será evaluado por un total de cinco expertos diferentes y para

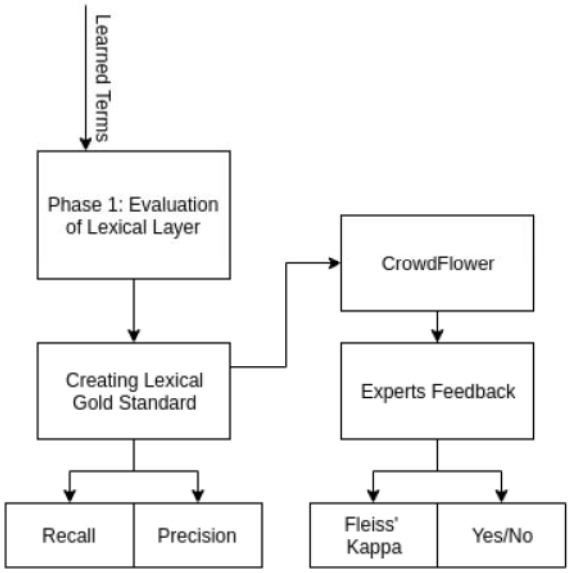

Figura 4.2: Diseño de la primera fase del proceso de evaluación

evaluar el acuerdo entre estos expertos se hará uso de una medida muy conocida en la estadística, como es Fleiss' Kappa [Fleiss and Cohen, 1973].

Fleiss' Kappa

La medida Fleiss' Kappa ofrece un valor estadístico sobre el nivel de acuerdo entre varios evaluadores. Se trata de una extensión del método Cohens' Kappa [Galton, 1892] que ofrece una medida sobre el acuerdo cuando únicamente existen dos evaluadores. La medida se describe en la ecuación 4.1 donde el denominador ofrece el grado de acuerdo entre los evaluadores sin que este se vea afectado por el azar y el numerador detalla el grado de acuerdo real entre los evaluadores.

$$k = \frac{\overline{p_a} - \overline{p_e}}{1 - \overline{p_e}} \tag{4.1}$$

En la ecuación 4.2 la variable p_a se obtiene a partir de la media de los valores de p_i (ecuación 4.2) para cada término y la variable p_e se obtiene a través de la ecuación 4.3.

$$\overline{p_a} = \frac{1}{N} \sum_{i=1}^{N} p_i \tag{4.2}$$

$$\overline{p_e} = \sum_{j=1}^{k} p_j^2 \tag{4.3}$$

La forma de calcular las variables p_i y p_j de las ecuaciones 4.2 y 4.3, se describe en las ecuaciones 4.4 y 4.5. La variable p_j (ecuación 4.4) calcula la proporción de asignaciones de cada una de las posibles respuestas j (por lo que, en la ecuación 4.3 se calcula la media de los cuadrados de estas proporciones), siendo N el número de términos que se han evaluado, n el número de expertos que han evaluado cada uno de los conceptos y $\sum_{i=1}^{N} n_{ij}$ la cantidad de jueces o expertos que han asignado el término i a la categoría (posible respuesta) j, que es fija. La variable p_i (ecuación 4.5) calcula el ratio de parejas de jueces que han llegado a un acuerdo en función de todas las posibles parejas de jueces (la media, utilizada en la ecuación general, se calcula en la ecuación 4.2). De la misma manera que en la ecuación anterior, n hace referencia al número de expertos que han evaluado los conceptos, k es el número de posibles respuestas o categorías y el término $\sum_{j=1}^{k} n_{ij}^2$ refleja el cuadrado de la cantidad de asignaciones para un término fijo en cada categoría.

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij} \tag{4.4}$$

$$p_i = \frac{1}{n(n-1)} \left[\left(\sum_{j=1}^{k} n_{ij}^2 \right) - (n) \right] \tag{4.5}$$

El valor final, k, se encontrará entre el rango $[-1, 1]$, siendo 1 un acuerdo total entre los evaluadores y -1 la inexistencia de acuerdo. Si el valor de k se encontrase por debajo o próximo al 0, el experimento no se podría llevar a cabo ya este valor reflejaría la imposibilidad de realizar una evaluación a partir del conjunto de términos utilizado.

En la Figura 4.3 se puede observar un ejemplo simple sobre el cálculo de esta medida. En este caso particular, el número de de términos a evaluar era de 17, las posibles categorías 2 (si o no), y el número de evaluadores por término 5. Después de realizar los cálculos, si se analiza el resultado final, se puede asegurar que el nivel de acuerdo entre los evaluadores no suficiente como para hacer uso de los resultados obtenidos a partir de las respuestas de los expertos para extraer medidas como precisión y recall, por lo que en este ejemplo, no se podría realizar la evaluación y habría que realizar cambios sobre el diseño del experimento, el conjunto de datos u otros aspectos del proceso.

4.1.2. Método de evaluación para la capa léxica

Una vez que se ha completado el proceso de creación del *Gold Standard* de términos se puede llevar a cabo la evaluación de la capa léxica. Para ello, se hará uso de las técnicas propuestas por [Sabou et al., 2005, Dellschaft and Staab, 2008] cuando se desea realizar una evaluación basada en *Gold Standard*. En este caso en

	A	B	C	D	E	F	G
1	**Term**	**Yes**	**No**	**Pi**			
2	graphics	5	0	1			
3	results	2	3	0.4		**N**	17
4	images	5	0	1		**k**	2
5	transactions	2	3	0.4		**n**	5
6	points	4	1	0.6			
7	models	4	1	0.6			
8	methods	3	2	0.4		**pa**	0.5764705882
9	data	3	2	0.4		**pe**	0.5753633218
10	objects	3	2	0.4			
11	values	2	3	0.4		k	0.00260756193
12	pixel	5	0	1			
13	constraints	3	2	0.4			
14	vertices	4	1	0.6			
15	techniques	2	3	0.4			
16	vertex	4	1	0.6			
17	regions	4	1	0.6			
18	parameters	4	1	0.6			
19	**Pj**	0.6941176471	0.3058823529				

Figura 4.3: Ejemplo del cálculo de Fleiss' Kappa

particular, se hará uso de las medidas de *recall, precision y F-Measure* detalladas en las ecuaciones 2.1, 2.2 y 2.3. En este apartado de la evaluación se le otorgará a la medida de *recall* una gran importancia, ya que de, de esta manera, se conocerá el porcentaje de conocimiento que se ha conseguido representar respecto al conjunto de conocimiento total que se debía representar, dejando en un segundo plano a la precisión, ya que dicha medida se incluirá en el proceso de evaluación de la capa taxonómica.

4.2. Fase 2: Construcción de relaciones taxonómicas

A lo largo de esta fase se construirá un *Gold Standard* de relaciones taxonómicas. Durante este proceso se hará uso de los términos aprendidos de forma automática, que hayan sido evaluados por expertos del dominio de *Computer Graphics* a lo largo de la fase anterior, y, además, que se haya probado que realmente son términos relevantes en dicho dominio. Es decir, todos los términos que hayan obtenido más votos positivos que negativos en relación a su pertenencia al dominio se considerarán como relevantes y formarán parte del *Gold Standard* de la capa léxica. Por ejemplo, si observamos los términos de la Figura 4.3, formarían parte del *Gold Standard*: *graphics, images, points, models, methods, pixel,* etc.

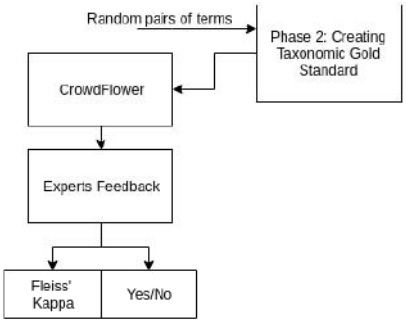

Figura 4.4: Diseño de la segunda fase del proceso de evaluación

A continuación comenzaría el proceso de construcción del modelo estándar. El método de creación será muy similar al descrito en la primera fase. En este caso, los expertos, en vez de evaluar si un término pertenece o no a un dominio, tendrán que decidir si una relación de tipo "*subclass of*" entre dos términos tendría sentido en una ontología que representase dicho dominio. Ese par de términos se escogerán de forma aleatoria a partir del conjunto de términos que forma el *Gold Standard* léxico. La construcción de este modelo estándar se realiza de esta forma con la intención de que sea lo más sencillo posible para los expertos, intentando reducir también el coste en tiempo. Si se continúa con el ejemplo de la Figura 4.3, se mostraría a los expertos preguntas del tipo: "¿Se cumpliría la relación *graphics is subclass of images* en un proceso de creación de una ontología en el dominio de los gráficos por computador?", a la que los expertos deberían responder únicamente de forma positiva o negativa. Al final de esta fase se obtendrá un conjunto de relaciones taxonómicas relevantes en el dominio que formarán el *Gold Standard* taxonómico, el cuál es necesario para realizar la evaluación de esta capa. El diseño de esta fase se puede observar en la Figura 4.4.

4.3. Fase 3: Evaluación de la capa taxonómica

A partir del *Gold standard* construido durante la fase dos y haciendo uso de las relaciones taxonómicas obtenidas durante el proceso de aprendizaje se puede dar comienzo a la evaluación de la capa taxonómica. De la misma manera que se describe en [Dellschaft and Staab, 2008], el proceso se dividirá en dos subprocesos, la evaluación a nivel local y la evaluación a nivel global. Todas las ecuaciones y técnicas de las que se hará uso durante el proceso se encuentran descritas dentro del apartado del *Taxonomic Framework* de [Dellschaft and Staab, 2008]. La propuesta de esta última fase de evaluación es, realmente, la parte innovadora del proceso, ya que la mayoría de ideas, sistemas o procedimientos no incluyen la evaluación

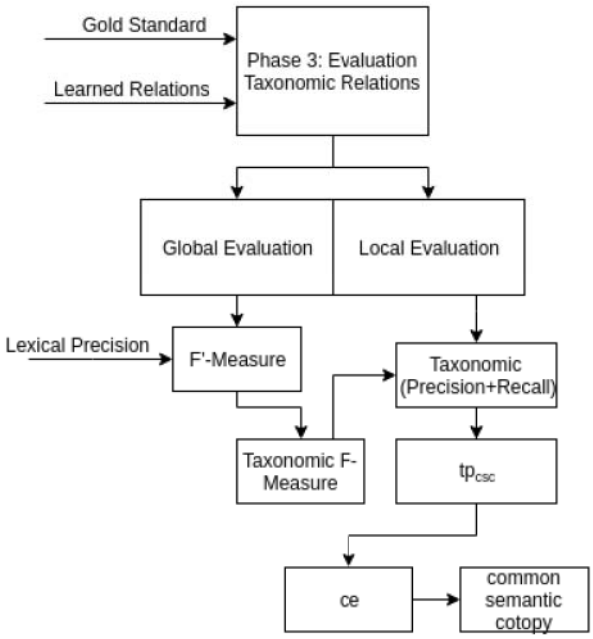

Figura 4.5: Diseño de la tercera fase del proceso de evaluación

de la capa taxonómica. Aun así, realizar dicha fase implica aceptar una serie de riesgos, ya que si la dificultad de los cálculos para ontologías simples es alta, en el momento en el que se desee evaluar ontologías más complejas, la dificultad de la fase aumentará de manera considerable. El diseño de esta sección de la evaluación se puede observar gráficamente en la Figura 4.5.

4.3.1. Evaluación global

La medida global final será la F'-Measure, descrita en la ecuación 2.12. Se escoge está función ya que calcula la media armónica entre el recall obtenido durante la primera fase del proceso, en la evaluación de la capa léxica (ecuación 2.2), y la *Taxonomic F-Measure* (ecuación 2.11), pudiendo así hacer uso de medidas que no se ven afectadas por la precisión en el apartado de evaluación local. La función *Taxonomic F-Measure* calcula la media armónica entre la precisión taxonómica (ecuación 2.9) y el recall taxonómico (ecuación 2.10), dando un valor de la relación entre estas medidas.

4.3.2. Evaluación local

Para las medidas locales, se hará uso de la precisión taxonómica y el recall taxonómico. Como se observa en la ecuación 2.10, la relación entre las dos medidas

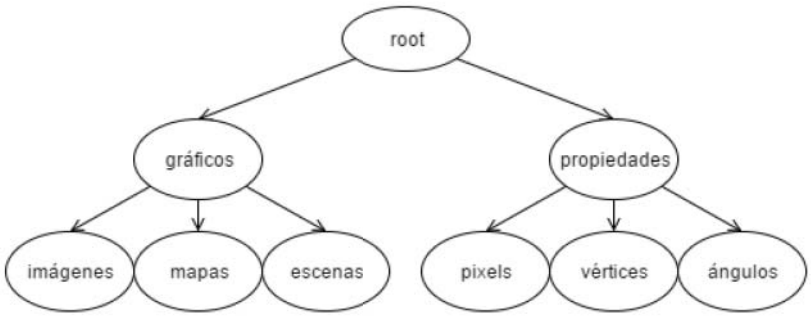

Figura 4.6: Ejemplo de jerarquía taxonómica estándar (Or)

es trivial, por lo que en la explicación se tendrá en cuenta, únicamente, la precisión taxonómica.

Como se desea que la precisión de la capa léxica no afecte a estas medidas con el fin de obtener resultados alterados, se hará uso de la ecuación 2.9, que define la precisión taxonómica en donde la variable ce, derivada de la ecuación 2.5 (tp_{csc}) de la que hace uso la TP_{csc}, toma el valor de la función 2.7, es decir, hace uso del *common semantic cotopy*. Esta ecuación, sólo tiene en cuenta los términos de la ontología aprendida que también se encuentran en la ontología de referencia o *Gold Standard*.

Un ejemplo muy sencillo sobre el funcionamiento de la evaluación de la tercera fase en el dominio de los gráficos por computador se podría realizar haciendo uso de los modelos de las Figuras 4.6 (jerarquía de referencia) y la 4.7 (jerarquía aprendida). A partir de estas jerarquías se debe calcular el *common semantic cotopy* como se refleja en la Tabla 4.1: para cada uno de los términos de cada una de las jerarquías se incluirán los subtérminos y supertérminos que se encuentren en ambas jerarquías. Por ejemplo, para el caso del término propiedades, el *common semantic cotopy* de la jerarquía de referencia serán root y vértices y para la jerarquía aprendida no habrá ningún término asociado ya que el término propiedades no aparece en dicha jerarquía.

A continuación se calcula la precisión y el recall taxonómico global haciendo uso de las ecuaciones 2.9 y 2.10, el sumatorio de la precisión local taxonómica partido entre el número de términos en común entre la jerarquía de referencia y la aprendida. En este caso la precisión y el recall taxonómico global obtienen el mismo valor: 5/6. A partir de este valor ya se puede calcular la medida global, F-Measure, ecuación 2.11. El valor de esta medida en el ejemplo es de 83,33 %. Para finalizar, se debe calcular primero el recall de la capa léxica: 5/9, y a continuación se obtiene el valor final a través de la medida F'-Measure (ecuación 2.12): 66,67 %.

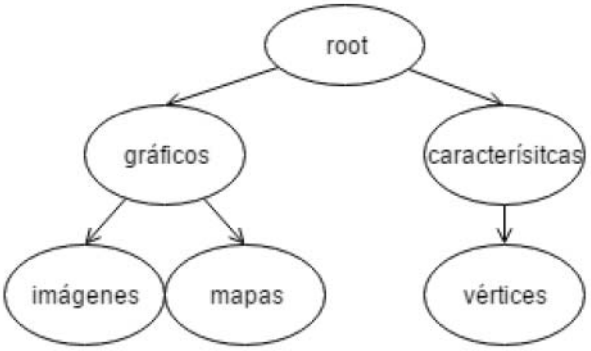

Figura 4.7: Ejemplo de jerarquía aprendida automáticamente (Oc)

c	**csc(c,Or,Oc)**	**csc(c,Oc,Or)**
root	{gráficos, imágenes, mapas, vértices}	{gráficos, imágenes, mapas, vértices}
gráficos	{root, imágenes, mapas}	{root, imágenes, mapas}
propiedades	{root, vértices}	-
características	-	{root, vértices}
imágenes	{gráficos, root}	{gráficos, root}
escenas	{gráficos, root}	-
mapas	{gráficos, root}	{gráficos, root}
pixels	{root}	-
vértices	{root}	{root}
ángulos	{root}	-

Tabla 4.1: Medición del *common semantic cotopy*

Capítulo 5

Implementación y experimentación

La implementación y experimentación se ha llevado a cabo en función de los pasos definidos en el capítulo 4. Como se ha comentado en capítulos anteriores, al hacer uso de expertos de dominio para la creación de *Gold Standard*, el tiempo necesario para la implementación es alto, por lo que en este caso, únicamente se han podido obtener resultados sobre la fase uno del experimento: *recall* y precisión de la capa léxica y el Fleiss' Kappa obtenido. Como se describe en el capítulo 6, una de las líneas de trabajo futura es la de realizar una evaluación completa con el fin de comprobar que el proceso diseñado ofrece resultados significativos.

A lo largo de este capítulo se describirá las características mas relevantes del dominio y los expertos del mismo en el que se ha centrado el experimento, los gráficos por computador, las herramientas utilizadas a lo largo de la experimentación, y finalmente, se mostrarán y analizarán los resultados obtenidos durante la ejecución de la evaluación de la capa léxica.

5.1. El dominio de "*Computer Graphics*"

El trabajo de fin de máster que se presenta se encuentra en el contexto del proyecto europeo *DrInventor* que se encuentra centrado en el dominio de los gráficos por computador. Es por ello, que el método propuesto debe hacer uso de un corpus de documentos afines a dicho campo para ejecutar la evaluación. El corpus de artículos anotados utilizado se encuentra almacenado en el siguiente enlace[1].

Como se ha comentado a lo largo del trabajo, es imprescindible la ayuda de expertos de dominio para poder llevar a cabo la creación de los modelos ontológicos estándar por lo que, en este caso en particular, los expertos pertenecerán al dominio

[1] `http://sempub.taln.upf.edu/dricorpus`

de los gráficos por computador. Cabe destacar que durante el experimento, ninguno de los evaluadores sabrá que los términos han sido aprendidos de forma automática, simplemente creerán que están evaluando una serie de conceptos escogidos por ingenieros ontológicos con el fin de crear una ontología en el dominio, de esta manera se asegura la objetividad del experimento.

Por último, cada uno de los términos, que durante esta primera aproximación de evaluación, se tratarán únicamente de sustantivos, se evaluará un total de cinco veces, no teniendo porque ser siempre los mismos expertos los que evalúan los términos. Esto se puede llevar a cabo gracias al uso de la medida de nivel de acuerdo entre expertos definida en el capítulo 4, Fleiss' Kappa, que permite realizar la evaluación haciendo uso de múltiples expertos.

5.2. Uso de herramientas para la generación de *Gold Standards*

Para poder llevar a cabo el experimento detallado en el capítulo anterior es necesario hacer uso de una serie de herramientas que, conjuntamente, permitirán obtener los datos que se necesitan para poder realizar la evaluación. En general, se tratan de herramientas sencillas: se hará uso del correo electrónico para enviar el enlace del experimento a los expertos, se utilizará el formato CSV[2] para la descripción de los conjuntos de datos y una hoja de cálculo para obtener las medidas deseadas sobre la evaluación. Sin embargo, para llevar a cabo la creación de *Gold Standards* de términos y relaciones taxonómicas es necesario la utilización de una herramienta más compleja, en este caso se ha escogido una aplicación web que permite realizar todos los pasos requeridos, ClowdFlower.

5.2.1. CrowdFlower

CrowdFlower es una herramienta que permite muchos tipos de tareas (análisis de sentimientos, categorización de datos, validación de traducción) sobre tipos de datos como imágenes, texto, etc. Se trata de una herramienta web muy potente que actualmente está siendo usada por algunas de las empresas tecnológicas más importantes a nivel internacional.

A continuación se describen las principales tareas a través de la aplicación web para poder llevar a cabo la realización de la evaluación y la creación de los *Gold Standard*:

- **Carga de datos:** El primer paso que se debe realizar es la carga de los datos. CrowdFlower permite cargar datos desde diversos formatos de archivo (.csv,

[2] `https://es.wikipedia.org/wiki/CSV`

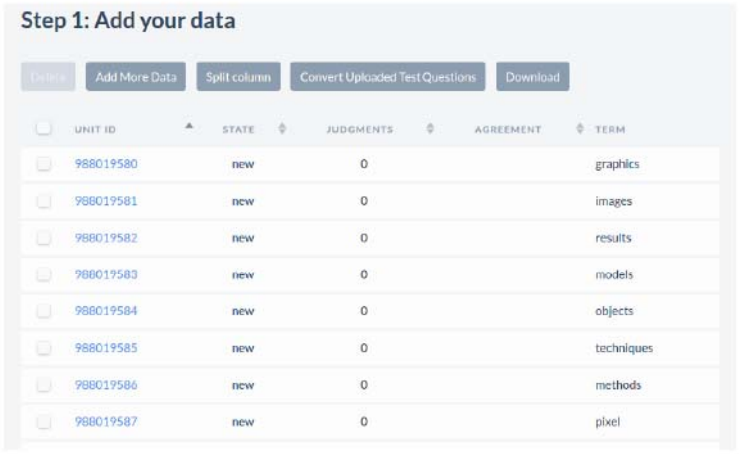

Figura 5.1: Edición del dataset a través de CrowdFlower

.tsv, .xls, .xlsx, .ods), en este caso, como se ha comentado anteriormente, se utilizará el formato CSV. El archivo únicamente contendrá una lista de los términos que se desean validar.

- **Edición del dataset:** Un paso intermedio entre el diseño del experimento y la carga de los datos es el de la edición del dataset subido a la aplicación. En este paso se pueden editar los datos subidos, descargar de nuevo el dataset entero, subir otro conjunto de datos o crear directamente preguntas de control. En la Figura 5.1 se puede observar una captura de pantalla de las funciones que ofrece este paso.

- **Diseño del experimento:** Es el paso más importante del experimento. En esta sección se introduce la pregunta que se desea formular a los expertos, en este caso, una pregunta muy simple sobre la pertenencia del término al dominio de los gráficos por computador. Además se definen las posibles respuestas a las preguntas, una descripción en lenguaje natural de las instrucciones necesarias para poder realizar correctamente el experimento y el título del mismo. En la Figura 5.2 se observa un ejemplo de la pantalla que se muestra en la aplicación web para realizar estas tareas.

- **Creación de preguntas de control:** El siguiente paso que se debe realizar es la creación de las preguntas de control con el fin de comprobar el conocimiento de los expertos sobre el dominio, la representación del conocimiento, etc. Como se ha comentado en apartados anteriores, en caso de que los evaluadores no contesten correctamente a estas preguntas, los resultados ofrecidos por los mismos no se tendrán en cuenta en el experimento. Para crear las preguntas de control, el creador del experimento debe responder a

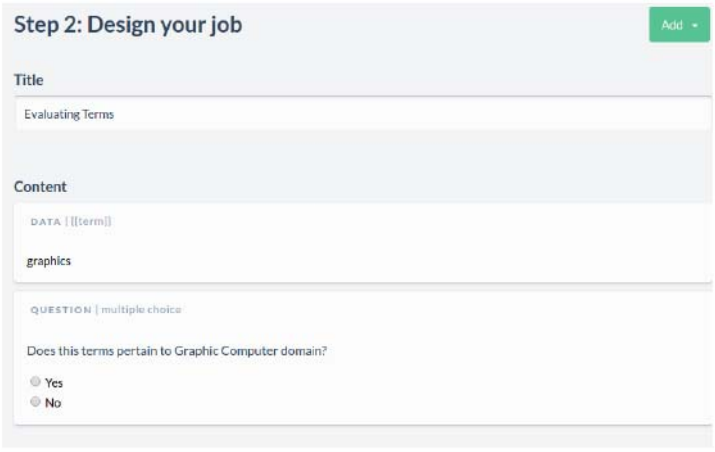

Figura 5.2: Diseño del experimento a través de CrowdFlower

la pregunta que se ha formulado en el apartado del diseño sobre una serie de datos en los que esté completamente seguro de su respuesta. Por ejemplo, en este caso particular, se podría responder con un "no" a la pregunta de si el concepto "table" pertenece al dominio de los gráficos por computador y "si" en el caso de que el concepto a validar fuese "image".

- **Lanzamiento del experimento:** El último paso que se debe llevar a cabo es la formalización del experimento y su lanzamiento. Para ello, se debe detallar que cantidad se le pagará a cada uno de los evaluadores por cada pregunta que se responda. A continuación se debe lanzar el experimento definiendo si se desea que cualquier persona registrada en CrowdFlower como evaluador pueda acceder al experimento y responder las cuestiones o, como en este caso, que se necesitan expertos de un dominio, generar una URL con la que se accede al experimento de forma privada. En la Figura 5.3 se puede observar la pantalla que se le muestra a los evaluadores con algunas preguntas sobre el experimento.

5.3. Resultados

En este apartado se detallarán y analizarán los datos obtenidos después de realizar el experimento sobre la primera fase del método propuesto y los resultados sobre el nivel de acuerdo entre expertos. Como se ha comentado con anterioridad, la complejidad del método de evaluación únicamente ha permitido obtener resultados de la primera fase, dejando como una de las líneas de trabajo futuro la realización de una evaluación completa del sistema. El conjunto de datos, así como los todos

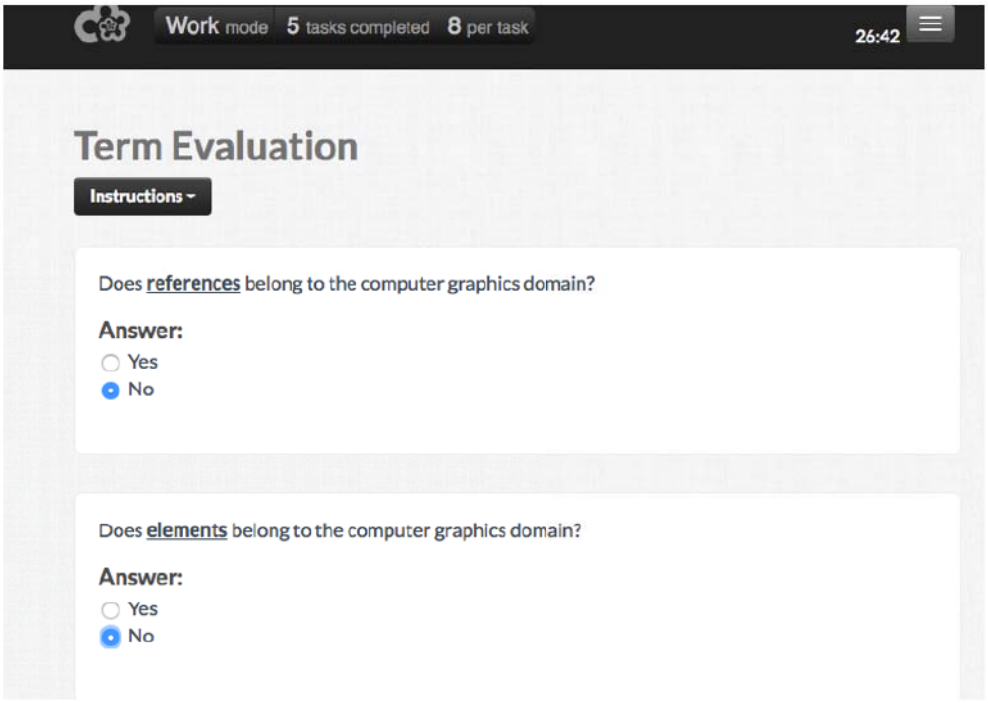

Figura 5.3: Captura del experimento realizado en CrowdFlower

los cálculos realizados para obtener los resultados que se muestran a continuación se pueden encontrar en [Chaves, 2016].

5.3.1. Evaluación fase 1

A partir del corpus utilizado en proceso de de aprendizaje se obtuvo un conjunto de 1788 sustantivos. Cada uno de estos términos tenia asociado una medida de relevancia, el *termhood*. Se trata de una medida estadística contrastada para el cálculo de términos relevantes a partir de un corpus de documentos y la manera de calcularla se describe en [Frantzi et al., 2000]. A partir del la lista de términos aprendidos se decidió realizar la evaluación de un subconjunto de dicha lista con el fin de comprobar el funcionamiento del proceso y obtener algunos resultados relevantes. Por esa razón, se estableció el umbral del *termhood* en un valor de 0,24 realizando la evaluación de los 104 términos más relevantes del conjunto aprendido. Un total de 17 expertos, todos relacionados con el campo de "*Computer Science*", respondieron a 520 preguntas, 5 validaciones por cada términos, y 12 preguntas de control. En la tabla 5.1 se muestran los resultados obtenidos para el recall, la precisión y el F-Measure (con $\beta = 1$) variando el umbral de relevancia de los términos (*termhood*). Además en la tabla 5.2 se detalla la precisión obtenida para

Termhood	Precision	Recall	F-Measure
0.23	67.3 %	100 %	80.45 %
0.27	68.7 %	47.1 %	55.9 %
0.32	81.8 %	25.7 %	39.1 %

Tabla 5.1: Medidas de evaluación de la capa léxica

Termhood	Precision5	Precision10	Precision20	Precision30	Precision40
0.23	60 %	70 %	80 %	80 %	75 %
0.32	60 %	70 %	80 %	81.8 %	81.81 %
0.42	60 %	75 %	75 %	75 %	75 %

Tabla 5.2: Medidas de precisión para de la capa léxica

los 5, 10, 20, 30 y 40 primeros términos mas relevantes con el fin de comprobar como esta medida se ve afectada al incluir o excluir más o menos términos.

Al analizar los resultados de la tabla 5.1 se puede observar que a medida que el umbral del *termhood* aumenta, la precisión también lo hace, mientras el recall y la F-Measure descienden. Esto quiere decir que a medida que disminuimos nuestro conjunto de términos relevantes aumentando el umbral del *termhood*, el conocimiento identificado correctamente en función de todo el conocimiento identificado aumenta, mientras que el conocimiento identificado en función del conjunto del conocimiento que se ha identificado a partir del corpus disminuye. Est

Por otra parte, si se analizan los resultados obtenidos de la tabla 5.2 se observa que la precisión para conjuntos de términos relevantes pequeños (5 o 10), esta medida no se ve muy alterada, pero a medida que incluyen más términos (40 o más) la precisión comienza a bajar debido a que se tienen en cuenta términos menos relevantes que generalmente, han sido evaluados como no pertenecientes al dominio de los gráficos por computador por los evaluadores.

5.3.2. Nivel de acuerdo entre expertos

Como se ha descrito en apartados anteriores, el nivel de acuerdo entre los expertos se calculará a partir de la medida Fleiss' Kappa. Como se observa en la tabla 5.3, debido a que el nivel de acuerdo era muy bajo entre los expertos se optó por implementar dos enfoques diferentes con el objetivo de aumentar dicho valor.

En el primer enfoque, la *black list*, se eliminaron los términos del dataset más comunes en el campo de la ciencia: *results, values, etc*, quedando un conjunto de 76 términos. En el segundo enfoque (*Agreement Filtered*), se optó por eliminar todos los términos que no tuviesen un p_i igual o superior a 0,6. De esta forma, se eliminaron los términos en los que los expertos dudaron más a la hora de evaluar,

	Normal	**Black List**	**Agreement Filtered**
k	0.16	0.18	0.38

Tabla 5.3: Nivel de acuerdo entre expertos

quedando un conjunto de 60 términos. Aunque el nivel de acuerdo entre expertos aumenta con este último enfoque, el valor de precisión se ve afectado negativamente, disminuyendo en gran medida. Los resultados obtenidos se pueden deber a un diseño del experimento algo incorrecto, ya que la pregunta realizada tenía un carácter abierto y no se obligaba a los expertos a tomar una postura a la hora de responder (conservadora u optimista). Para solucionar esta cuestión se debería volver a lanzar el experimento con una pregunta concreta que no hiciese dudar a los expertos.

Capítulo 6

Contribuciones y trabajo futuro

En este capítulo, se resumen las principales contribuciones que aporta este trabajo en el campo del *Ontology Learning*, y más específicamente en la evaluación de resultados de dicho campo. Además, se detallarán las principales líneas de trabajo futuro con el fin de que contribuyan a la mejora de las ideas presentadas durante este trabajo.

6.1. Contribuciones

Las contribuciones que se han realizado en el campo del *Ontology Learning* por parte de este trabajo se han centrado en el apartado de la evaluación de ontologías aprendidas de forma automática.

La principal contribución de esta tesis fin de máster ha sido la presentación de un método integrado de evaluación para ontologías aprendidas automáticamente. El objetivo de dicha aportación es la de ofrecer, a los desarrolladores de algoritmos de aprendizaje, un *framework* estándar para la evaluación de cualquier ontología de dominio. Con la intención de automatizar un proceso que suele ser tedioso, el sistema se ha construido en base a técnicas y métodos cuantitativos, que suelen ser más propensos a la automatización. No obstante, como se ha detallado en el capítulo 5, actualmente es necesaria la participación de expertos del dominio. La principal razón es que, aunque sólo sean necesarios para la creación de los modelos estándar, la automatización de este paso no permite asegurar que los modelos reflejen correctamente el conocimiento del corpus de documentos.

Cabe destacar que el método de evaluación presentado a lo largo de este trabajo formará parte del conjunto de aportaciones presentadas por el OEG[1] para el congreso EKAW2016[2]. La aportación enviada al congreso se encuentra detallada

[1] `http://www.oeg-upm.net/`
[2] `http://ekaw2016.cs.unibo.it/`

en el Apéndice A.

6.2. Trabajo Futuro

Durante la realización de este trabajo se han podido identificar un conjunto de líneas de trabajo futuras que podrían contribuir a la mejora de las propuestas realizadas completando así, las ideas descritas a lo largo de del documento. Estas líneas son:

- **Validación del método de evaluación propuesto:** Como se ha observado en el capítulo 5, debido a las restricciones de tiempo de la tesis final de máster, el proceso de experimentación ha sido muy corto. Cabe destacar, que dichos procesos, al necesitar a expertos humanos, se suelen extender bastante en el tiempo, por lo que la presentación de resultados completos en el trabajo era una tarea complicada. Es por ello que la línea de trabajo futura principal sería la de realizar un proceso de validación del sistema con el fin de comprobar si el método ofrece unos resultados adecuados.

- **Mejora del proceso de creación del *Gold Standard*:** A lo largo del proceso de evaluación de ontologías descrito en el capítulo 4, se puede observar que los pasos con más coste para el sistema, son los de la creación de los modelos ontológicos estándar para la evaluación de la capa léxica y la capa taxonómica. Es por ello, que una línea de trabajo futura sería la automatización sobre la creación de estos modelos. Podría ser un buen punto de partida las ideas desarrolladas en [Spyns, 2005] incluyendo tecnologías, algoritmos y métodos actuales de procesamiento de lenguaje natural. La propuesta final podría hacer uso de los expertos para evaluar, únicamente, el modelo estándar construido de forma automática, lo que disminuiría en gran medida el tiempo necesario para la realización de la evaluación consiguiendo así un equilibrio entre este recurso y la calidad del proceso.

- **Automatización del sistema de evaluación**: Otra de las cuestiones que se podría tratar como línea futura sería la de la creación de un sistema *ad-hoc*. En la propuesta actual, se hace uso de múltiples herramientas (email, Crowdflower, CSVs, hojas de cálculo) lo que provoca que el proceso sea tedioso, tanto para los evaluadores como para los desarrolladores del mismo. Por estas razones, la implementación y desarrollo de una aplicación web similar a Crowdflower pero con la posibilidad de visualización en el propio sistema de los resultados (tanto de la evaluación como del nivel de acuerdo entre expertos), la subida de varios *datasets* diferentes, etc. beneficiaría al sistema, mejorándolo y dotándolo de robustez.

- ***Ontology Matching***: A lo largo del estado del arte no se ha conseguido encontrar algún método que hiciese uso de métodos de *Ontology Matching* para realizar la evaluación. Debido a esto, y, a que el enfoque que proponen estás técnicas podría ser muy útil en el proceso de evaluación sustituyendo la complejidad de las medidas propuestas en [Dellschaft and Staab, 2008], se debería realizar un estudio e implementar un experimento con el fin de analizar como afectaría la introducción de métodos de *Ontology Matching* en el sistema.

- **Evaluación de la capa no-taxonómica:** Como última línea de trabajo futura, abordar la capa no-taxonómica de las ontologías sería importante para completar el proceso de evaluación. Aun así, tanto en el apartado de aprendizaje, como en el de evaluación, existe una complejidad muy alta debido a la naturaleza de las relaciones que, a día de hoy, dificulta en gran medida la posibilidad de incluir este apartado en el proceso.

Apéndice A

Artículo EKAW2016

Towards an Integrated Approach for Ontology Learning Evaluation

José Luis Redondo García, David Chaves Fraga, Carlos Badenes Olmedo, Óscar Corcho

Ontology Engineering Group, Universidad Politécnica de Madrid,
jlredondo@fi.upm.es, david.chaves.fraga@alumnos.upm.es,
{cbadenes, ocorcho}@fi.upm.es

Abstract. *Ontology Learning* algorithms are used to automatically generate ontologies, usually from unstructured resources. They are specially useful in particular domains where there are no mature or de-facto ontologies. This discipline is re-flourishing thanks to recent advances in information extraction techniques exploiting novel features and paradigms such as Word2Vec and Deep Learning. Ontology learning algorithms work over very heterogeneous and continuously evolving data sources, which makes it difficult to create a universal, scalable evaluation methodology that allows to quantitatively determine their adequacy for the generation of domain ontologies. In this position paper we present a unified approach for evaluating ontology learning algorithms, which takes into consideration different lexical and taxonomical aspects that are compared against a semi-automatically generated Gold Standard. This strategy minimises human intervention and promotes more replicable experimental setups. Through various use cases, we have proven how the proposed methodology can adapt to different scenario requirements while increasing the representativeness of the Gold Standard, therefore providing more meaningful insights about the candidate techniques.

Keywords: Ontology Learning, Evaluation, Natural Language Processing, Knowledge Representation

1 Introduction

Every second, huge amounts of data about any imaginable topic are being generated or captured, and therefore become ready to be exploited. A significant subset of this information is available in the form of natural language text, which needs to be interpreted. Unfortunately, the most advanced agents to perform such a task (we, the humans) are not capable of processing a significant fraction of this data in a reasonable amount of time.

Ontology Learning algorithms are used to automatically generate ontologies from a set of raw textual documents. Ontologies allow users to understand, from a single information unit, the kind of knowledge being represented by the underlying documents. They also help machines to better exploit data by contextualising and customising different operations over the items inside the corpus. Given the importance of such conceptualisation, different methodologies for evaluating the quality of the automatically generated ontologies are needed. Early approaches [6] in the research field tackled this evaluation by comparing the generated vocabularies against some well-stablished ontologies in the domain. Consequently, the objective is to build a model with the same level of generalisation and high formal restrictions than the reference ontology, which has been normally engineered by humans. Given that a domain ontology is already available, the learning process becomes less important and approaches tend to adopt a top-down paradigm (see left side of Figure 1) where data is matched to the model and not the other way around [2].

However, today's information extraction and knowledge representation scenarios are different. Innovative techniques to generate features from text, such as Word2Vec, or paradigms like Big Data or Deep Learning allow processing huge amounts of information with higher accuracy and therefore target a wider variety of domains. In this situation, the task of finding already-existing ontologies that sufficiently match the data being analysed becomes unfeasible in most of the cases. In addition, the knowledge about the domain is quickly evolving so models need to be able to capture the changeable context and react to new trends reshaping the conceptualisation over time.

Therefore, former ontology learning efforts relying on already-existing reference ontologies may not be enough anymore. Current approaches are converging towards the need of more flexible, lightweight, local models that can closely describe the corpus and are easily updatable. As depicted on the right side of Figure 1 they are generated solely from the data with no influence of previous established models, and aiming for a less strict degree of formalisation than what traditional ontology learning approaches wanted to capture.

To the best of our knowledge, there are not research efforts trying to *evaluate* this new kind of automatically generated models. In this

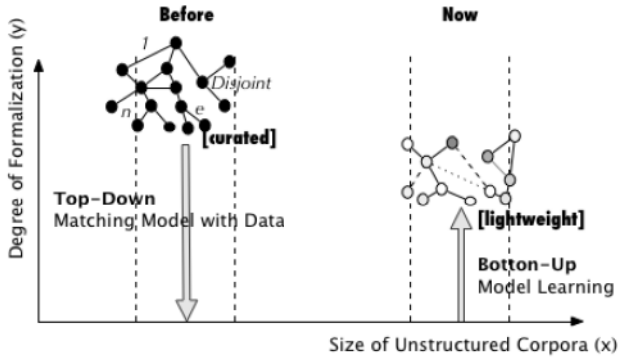

Fig. 1: Increasing importance of ontology learning evaluation in current scenarios

position paper we propose a new methodology to deal with the new requirements of the reemergent field of ontology learning, aiming to identify an incremental set of evaluation objetives and methods that can lead to more effective judgements about the quality of the learned model.

The remainder of the paper is organised as follows: Section 2 reviews the most prominent work in the field of ontology learning and its evaluation, Section 3 gathers the definitions that support the evaluation methodology presented in Section 4, and Section 5 explains how this methodology has been partially and implicitly applied over previous works, in order to highlight its potential.

2 Related Work

Ontology Learning is a wide discipline that considers a great variety of methods and techniques, some of them reported in [3]. In the particular case of ontology learning from text, some relevant approaches have been described in [24]. The methodology presented in this paper has been initially inspired by the paper by Dellschaft et al.[9], which already formalises an approach for evaluating ontology learning algorithms. The main difference with our work is that their evaluation is grounded on the existence of de-facto ontologies, so the utilised methods come down to the application of ontology alignment techniques. In similar work from the same authors[8] the use of a gold

standard is described as a very desirable method for ontology learning evaluation. We have revisited this idea for our methodology, in an attempt to make it more lightweight.

The ontology learning system Syndikate[14] implemented an incremental algorithm exploiting evidences and certain credibility hypotheses about concepts that are refined through a supervised classification method trained on related knowledge bases. The evaluation merged a comparison to an already-existing corpora from information technology with some manual assessments on the generated hypothesis, making this method hardly reproducible by others. Text2Onto [6] selected another well-known field (tourism), and compared their approach against an already available domain ontology for applying precision and recall measures on terms and relations. The set of unnamed, non-taxonomic relations were hand-coded into the very same ontology before. Hence, they incurred on significant costs derived from the human efforts made, what we try to alleviate in our proposal. The system at [21] presents an approach for learning ontologies in bioinformatics. In this case, the evaluation methods verified the quality of the generated ontologies by comparing them against reference ontologies which were not initially available, but were hand-built for the occasion.

More recent examples like the framework Galeon [17] also performs an evaluation over terms and hypotheses (relations) from a traditional ontology-matching oriented point of view, by comparing the learned ontology with reference ontologies in the domains of universities and economics. In [5] they use an interesting mechanism to build a synthetic dataset to compare with, however we argue if the absence of lightweight human intervention can still lead to quality Gold Standards. They also perform a criteria-based evaluation considering aspects like "Mean to Root" or"Mean to Parent". However we will opt for more functional measures, which are easier to interpret in isolation. Finally, CRCTOL [15] distinguishes between a "component level" and an "ontology level" during the evaluation. The former phase consists in a lexical comparison with a gold standard in the terrorism and sport domains in order to quantify the performance against other systems like Text2Onto. The latter performs an evaluation on the relations learned, using quantitative and qualitative methods, and including an analysis of the graph struc-

tural properties, comparison to WordNet, and expert rating. This is probably the most exhaustive evaluation methodology that can be found in the literature, but it is not properly formalised, the gold standard annotations were generated from scratch by humans, and human efforts are non reproducible. From the different alternatives studied here, we can observe how none of them has followed a well formalised, easily applicable methodology sufficiently aiming to reduce the high costs derived from human intervention.

3 The Ontology Learning Evaluation Task

In this first section we propose a set of definitions that will be useful to describe the objectives and design decisions in our ontology evaluation methodology.

3.1 A Definition of Ontology for Automatic Learning Tasks

Below we formalise the concept of an ontology that aims to better match the specific requirements of the ontology learning domain. In contrast with much more complex formalisations of ontologies [10], this research topic relies on less constrained ontologies that are automatically built following principles such as incremental generation (the model evolves as more items in the corpora are processed) or flexibility (knowledge can change when more items are added into the corpora). According to this, we introduce the definition of a flatten ontology as a triple composed by the sets, $\boldsymbol{W}$, $\boldsymbol{R}$, $\boldsymbol{P}$:

$$\mathcal{O} = \{\boldsymbol{W}, [\boldsymbol{R}], [\boldsymbol{P}]\} \tag{1}$$

Definition 1. *Flatten Ontology Simplified representation of an ontology considering three different sets: the terms $\boldsymbol{W}$, the set of relationships between terms $\boldsymbol{R}$, and the global metadata properties $\boldsymbol{P}$. It focuses in the functional dimension of the ontology, leaving aside some structural details that are not relevant in recent automatic learning tasks.*

The first set of terms in $\boldsymbol{W}$ is defined as $\forall w \in \boldsymbol{W}, w \in \mathbb{S}$, being $\mathbb{S}$ the set of all strings generated as a combination of letters in our

alphabet. $\boldsymbol{R}$ is the set of all the relations established between pairs of terms w_a and w_b, formalised as a triple $\forall r \in \boldsymbol{R}, r = \{w_a, w_b, c\}$, where $c \in \mathbb{S}$ specifies the kind of connection established between w_a and w_b. The last set $\boldsymbol{P}$ defines the different metadata properties that sometimes are further characterising certain ontologies: general description, list of keywords, etc. Each property $p \in \boldsymbol{P}$ is a pair of name n and a string value v. The set of relations $\boldsymbol{R}$ and metadata properties $\boldsymbol{P}$ can be empty in order to keep the formalisation flexible enough for different learning tasks.

3.2 Ontology Learning Evaluation Objectives

Trying to capture the knowledge about a particular domain or subdomain can result in heavyweight formalisations with many axioms or restrictions, which may be expressed in formal languages like OWL[1]. However not every task leveraging on ontologies requires this degree of specificity. On the one hand, real world applications rarely need very complex representation models, because they are too complicated for expert users and developers. On the other hand, ontology learning techniques employ many state-of-the-art approaches that are still far from being able to deal with details such as cardinalities or universal quantifications.

In order to better address this complexity, in this research work we introduce the notion of *Evaluation Objectives* ω:

Definition 2. *Ontology Learning Evaluation Objective. Given an ontology $\mathcal{O}$, an evaluation objective ω is a particular subset of $\mathcal{O}$ that we aim at evaluating. Given the formulation of a Flatten Ontology previously presented in Equation 1, those evaluation objectives can therefore be the lexical $\boldsymbol{W}$, taxonomical $\boldsymbol{R}$, or general metadata $\boldsymbol{P}$, layers or all their possible subsets and combinations.*

Depending on the objectives established on each ontology learning initiative, those evaluation objectives can vary from the more basic, lexical-oriented goals, to others putting emphasis on the relations between the identified concepts.

[1] `https://www.w3.org/TR/2012/REC-owl2-syntax-20121211/`

3.3 Ontology Learning Evaluation Methods

Having considered a set of evaluation objectives $\omega_1, \omega_2, ..., \omega_n$ we need different methods to determine how good those features were learnt. Hence we introduce the definition of *Evaluation Method* $\mathcal{F}$:

Definition 3. *Evaluation Method. Given an automatically generated ontology* $\mathcal{O}$*, and an evaluation objective* ω*, an Evaluation Method is a function* $\mathcal{F} : \omega \rightarrow \mathbb{R}$ *that expresses into a unified score the degree of success of the learning algorithm for automatically generating* ω*.*

We consider as the ideal output of the learning approach, what an unbiased set of expert human annotators with infinite amount of time would have chosen as candidates after analysing the entire corpus.

Quality Criteria for Evaluation Methods Besides the good principles already reported in the literature for measuring different information tasks and systems [18], below we identify a set of desirable characteristics that ontology learning methods should follow in order to better serve their purposes. They are grounded on the very particular needs of the ontology learning task, which demands very changeable and easily formalisable knowledge models, generated by iterative and train-based algorithms that may require the method $\mathcal{F}$ to be executed multiple times.

- Cost-effective. Algorithms learning patterns from data are sometimes based on supervised approaches [22] that usually require the evaluation to be performed multiple times over a particular subset of the data, so the method has to be affordable to be executed in terms of time and resources.
- Reproducible. The method $\mathcal{F}$ has to be easy replicable not only by a specific learning approach but also by other systems being developed a posteriori.
- Extensible. The function $\mathcal{F}$ should be relatively easy to update in case that new considerations or observed facts are later considered, in order to offer the maximum level of trustiness.

We decide to systematically discard qualitative-oriented methods, which are much more subjective, difficult to define, and complicated to interpret.

Types of Evaluation Methods Evaluation methods $\mathcal{F}$ can be implemented following very different philosophies and techniques. In the list below we introduce some of the most relevant ones, inspired on a similar classification in [9].

- Task-based ($\mathcal{F}_{Task}$). This kind of evaluation methods try to measure how much a system improves in performing a certain task when an ontology is integrated into its workflow [19]. The problem with this perspective is that the methods are so specific that it is complicated to find well-suited measures to be applied. Also, it is influenced by implicit factors that make harder to solely attribute the improvements to the use of a certain ontology.
- Criteria-based approach ($\mathcal{F}_{Criteria}$). In this case certain expected patterns, properties and rules are set beforehand and checked over the results of the algorithm being tested [13]. This kind of techniques are very appropriate for programatic evaluations, but they are sometimes difficult to interpret and justify so they have to be further supported by other evaluation methods.
- Corpus-Based ($\mathcal{F}_{Corpus}$). They are good in measuring functional aspects, easy to automate and reproducible by third parties. However they assume that the corpus used as ground truth is representative enough of the domain, so the process of generating such dataset can end up being significantly tedious and exhaustive. Some examples of this methods are described in [8].
- Assessment ($\mathcal{F}_{Assess}$). Experts in the domain or potential consumers of the results go through the output of an algorithm to judge on their validity. This is the most intuitive way of implementing an evaluation method, at the risk of not being able to define a clear set of guidelines that align all annotators into a well defined task and the difficulty to recreate certain conditions to fairly compare the current approach with other systems.

The Cost of an Evaluation Method Going deeper into the criteria introduced above, one of the most important aspects to be taken into account when selecting one particular evaluation method against others is the cost $\mathcal{C}(\mathcal{F}) : \mathcal{F} \rightarrow \mathbb{R}$ of executing it. It may happen that a method is highly reliable in determining the quality of an ontology learning algorithm, but the cost of execution is so high

that researchers will be discouraged to use it, moving to less desirable practices instead. Without aiming to be exhaustive and just to emphasise the existence of different temporal and resource-based costs associated to each evaluation method $\mathcal{F}$, we identify two different kinds of costs: the one associated to techniques involving humans in the evaluation, $\mathcal{C}(\mathcal{F}_{Human})$, and the one derived from the automatic execution of an algorithm $\mathcal{C}(\mathcal{F}_{Automatic})$. For the sake of simplicity, we consider that the human costs associated to crowdsourcing campaigns (higher number of annotators, average/low knowledge about the domain) are pretty much the same than the ones relying on experts (lower number of annotators, deep knowledge about the domain). We also differentiate between a creational process where annotations are being generated from scratch by humans, and a validation process where some already existing learning results are just being judged: costs, $\mathcal{C}(\mathcal{F}_{Create})$ and $\mathcal{C}(\mathcal{F}_{Automatic})$. We establish two main premises that will influence our later decisions in the ontology learning evaluation methodology described in section 4:

$$\forall \mathcal{F}_i, \mathcal{C}(\mathcal{F}_{Automatic}) \ll \mathcal{C}(\mathcal{F}_{Human}) \tag{2a}$$

$$\forall \mathcal{F}_i, \mathcal{C}(\mathcal{F}_{Assess}) \leq \mathcal{C}(\mathcal{F}_{Create}) \tag{2b}$$

3.4 Evaluation Methododology

Having described the concepts of ontology learning evaluation objective ω and ontology learning evaluation method $\mathcal{F}$, we finalise by formalising the notion of an evaluation methodology $\mathcal{M}$. Given that, for each ω_i we need a evaluation method operating over it, we also introduce the concept of dimension $d = (\omega_i, \mathcal{F})$ as a way to make explicit this pair.

Definition 4. *Evaluation Methodology. An evaluation methodology is a list of evaluation dimensions d, formalised as a pair* $(\omega_i, \mathcal{F})$ *where for each evaluation objetive* ω_i *identified there is a corresponding* $\mathcal{F}_i$ *that allows evaluating it.*

$$\mathcal{M} = \{d_1, d_2, d_3, ...d_n\} \tag{3}$$

4 An Integrated Ontology Evaluation Approach

In this section we present our unified vision on the evaluation of ontology learning algorithms. We intend to cover the most relevant use cases being considered in the field, from the more relaxed thesaurus-oriented approaches that only concern about finding sets of terms, to the more tightly constrained efforts that consider also relations between those terms, or even global indicators describing the ontology as a whole. In addition and as stated in the introduction, we consider an ontology learning scenario where the diversity and changeability of the datasets make it difficult to find a good ontology matching the underlying data.

In this position paper we introduce a new ontology learning evaluation methodology $M_{OntoLearn}$ that aims to normalise the evaluation procedure along the different research efforts in the field. Our contribution is twofold:

1. We propose three evaluation dimensions d_w, d_r and d_p targeting different aspects represented in our definition of *flatten ontology* introduced in section 3. This way we integrate into a single methodology different needs from the ontology learning community.
2. We propose an evaluation method called Hybrid-GS (denoted as F_{Hybrid}) to be applied on at least the two first dimensions d_w and d_r. This methodology extends the corpus-based methods already applied in the literature [8] [26], through the application of principles to ensure good coverage and similar quality to pure human-driven approaches.

Therefore, and taking advantage of the definitions in section 3 and the notion of flatten ontology $\mathcal{O}$ we define our methodology as follows:

$$\mathcal{M}_{OntoLearn} = \{(\omega_W, F_{Hybrid}), (\omega_R, F_{Hybrid}), (\omega_M, F)\} \qquad (4)$$

Fig 2 depicts the whole integrated evaluation process, presenting the 3 different dimensions proposed $M_{OntoLearn}$ and how the F_{Hybrid} method can be used to evaluate at least levels 1 and 2.

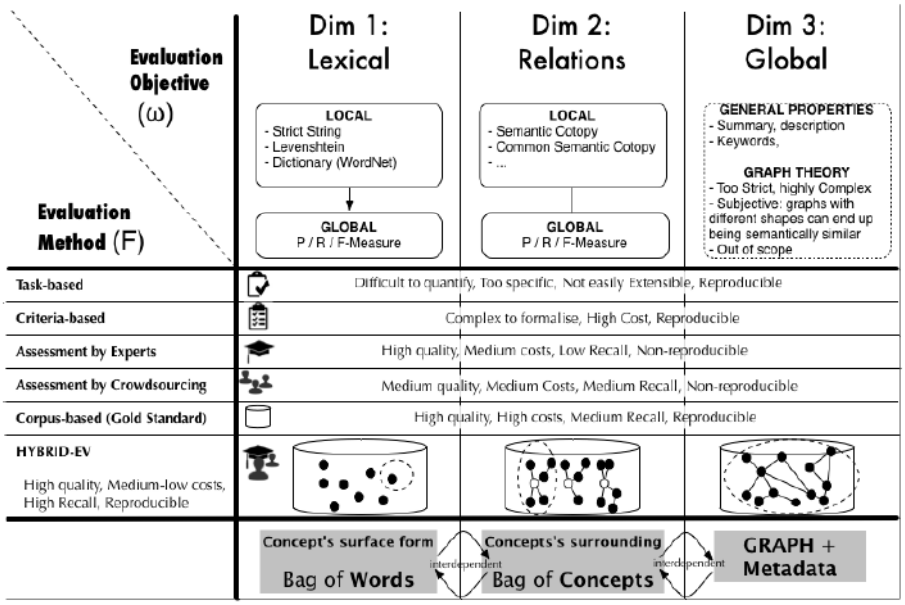

Fig. 2: General view of the Evaluation Methodology

We observe how the columns in the figure correspond to the three research objetives on which our methodology is based: terms, relations, and general properties. Most of the approaches in the literature target at least the first dimension since they focus on identifying a set of terms, while remaining agnostic to connections between them. The second dimension tries to determine the adequacy of the taxonomic and non-taxonomic relations established between terms. The third and last dimension considers the ontology as a whole, sometimes using graph-based measures, so that it can be complicated to evaluate and very difficult to interpret. It can also involve different global properties that are easier to compare, but hardly reused by ontology learning tasks. The rows in figure 2 represent the different evaluation methods $\mathcal{F}$ that may be used for evaluating the aforementioned research objectives. The reasons that led us to propose a new method F_{Hybrid} were mainly: 1) task-based approaches are difficult to quantify and specific, and the improvements in the scores cannot be straightforward linked to an increase in quality of the learned ontology, 2) criteria-based methods are complex to formalise, jus-

tify and maintain, 3) assessment from experts and crowdsourcing campaigns provide good quality but normally low recall and reproducibility, and 4) traditional gold-standard approaches offer good quality annotations but incur in significant costs and do not ensure high coverage.

4.1 The Method Hybrid-GS

We introduce a novel evaluation method that is inspired by the corpus-based evaluation methods [8], but introduces various advantages with the objective of reducing the cost of generating a gold standard while maintaining an adequate quality in the annotations generated. The so called F_{Hybrid} method aims to leverage on automatic annotation techniques as much as possible, while still performing a human assessment on top that ideally is more lightweight than creational processes (see equation 2b) and performed only once. This way, we get the best of the three worlds: corpus-based methods that are highly reproducible, automatic approaches that can be executed at a lower cost (see equation 2a) and bring higher coverage by quickly processing large sets of documents in short time, and the precision of human assessments, especially in user oriented tasks.

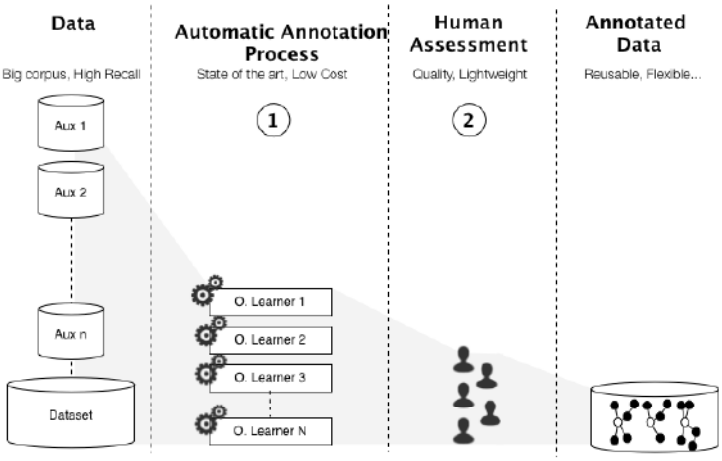

Fig. 3: Data selection flow in F_{Hybrid} evaluation method

The method F_{Hybrid} is composed by two selection phases, labeled in Figure 3 as *Automatic Process* and *Human Assessment*. In the first one, different state-of-the-art automatic learning techniques (ideally more than one and never the one being benchmarked) are executed in parallel to produce a first set of annotations. The second phase is performed by humans and consists on validating the integrated results from the ontology learners. Thanks to the use of automatic techniques in the first step, the original corpora being annotated can be accompanied with other relevant datasets, significantly increasing the coverage of the results. To better summarise those particularities and complement the hypothesis in equation 2, we formulate the following inequalities:

$$\mathcal{C}(\mathcal{F}_{Human}) \ll \mathcal{C}(\mathcal{F}_{Hybrid}) \leq \mathcal{C}(\mathcal{F}_{Auto}) \tag{5a}$$

$$Prec(\mathcal{F}_{Human}) \approx Prec(\mathcal{F}_{Hybrid}), Rec(\mathcal{F}_{Auto}) \approx Rec(\mathcal{F}_{Hybrid}) \tag{5b}$$

4.2 Dimension 1: Lexical

This dimension (ω_W, F_{Hybrid}) has as ultimate objetive ω_W to evaluate the lexical aspect of the ontology being built. We focus on verifying whether the set of terms t automatically generated correspond to what an unbiased set of expert annotators with infinite time would have chosen after analysing the entire corpus.

Therefore the evaluation objetives will come under the form of a bag of terms $\{t_1, t_2, ...t_n\}$ if the order does not matter, a list $(t_1, t_2, ...t_n)$ if they are ranked in importance, or similar aggregations. For evaluating the quality of those results one solution is to apply traditional Precision and Recall methods over the whole set of terms automatically extracted (labeled as "Global" measures in Figure 2), by comparing them with the Gold Standard ideally created by methods F_{Hybrid}, which contains the set of W as specified in the definition of flatten ontology. In cases where order matters, other measures from information retrieval can be used, such as *Mean Average Precision* (MAP) or *Normalised Discounted Cumulative Gain* (NDCG) [7]. We would like to emphasise that some of those measures are too focused in the performance at the top positions of the automatically retrieved list, therefore neglecting the big picture of the

domain vocabulary that we are trying to generate programatically. Other measures more oriented to coverage are normally preferred since today's data exploitation tasks tend to prioritise representativeness of the learned information unit against very high precision (see special measures such as *Compactness* reported in [12])

To obtain this global score we need a local similarity measure to operate between pairs of items from both the ontology learning results and the Gold Standard. We consider two kinds of distances, 1) the ones relying solely on the terms' surface form (Strict Distance and Relative String Distances[2] such as the Jaro-Winkler distance), and distances leveraging also on external knowledge sources for further improving the comparison, such as WordNet [1] or DBpedia [16].

4.3 Dimension 2: Relations

This dimension (ω_R, F_{Hybrid}) has as ultimate objetive to check on the potential relations established between the previously identified terms (ω_R), whether they are taxonomic (hyponymy and hypernymy) or not (thematic). Unlike other approaches that try to decouple the dependency between the lexical and the relational levels [9], we acknowledge this singularity and assume it as an inherent characteristic of the ontology constituents. To capture this notion of relations, we focus on verifying what we call concept c, defined as a subset of relations $r_c \in R_c$ and terms $w_c \in W_C$ such that there exists a set of n relations in and $n-1$ terms that connect w_c with the seed term t_c, being n the maximum depth considered. Similar approaches leveraging on local neighbourhoods have been introduced in other research work like [25].

The evaluation objetive will come under the form of a bag of concepts $\{c_1, c_2, ...c_n\}$, or similar aggregations. As in the previous dimension, we will be mainly interested in applying traditional Precision and Recall methods over the whole set of concepts by comparing them with a Gold Standard, ideally created by following the method F_{Hybrid} but now adding also relations R to the list of terms W. For implementing the local measure, there are different distances $D(c_a, c_b) \rightarrow \mathbb{R}$ in the literature that rely on the surrounding terms and relations [25]. Here we highlight two different measures called

[2] `https://en.wikipedia.org/wiki/String_metric#List_of_string_metrics`

semantic cotopy and *common semantic cotopy*, as described in [8]. It is worth noting that all the terms and relations between concepts belong to the ontology being learnt and not to external sources like WordNet or DBpedia, which fall into the lexical dimension.

4.4 Dimension 3: Global

The last considered dimension looks at the big picture of the ontology by analysing it as a whole. The objective ω_P takes into account global properties of the ontology such as topics, keywords, or summaries, but it can still highly leverage on the set of terms W and properties P, revealing once again how dimensions are incremental in complexity but interconnected with previous levels. These kinds of methods are far more complex to apply, and this is why we haven't specified a particular kind of function in the definition of our methodology $\mathcal{M}_{OntoLearn}$. However F_{Hybrid} methods can be equally applied, finally getting to construct an entire *flatten ontology* that can be used as a Gold Standard. We have identified three kinds of global-oriented evaluation methods:

1. As the net of terms and relations are shaping up a graph, we can check on its isomorphism against the reference ontology. In this particular case, we could re-use the same Gold Standard generated for evaluations considering dimension 2 that already includes nodes and connections. Examples of such similarity functions can be found at [11] where the authors rely on a graph edit distance. The problem of this kind of evaluations is that the morphological differences between the two graphs do not necessarily correlate with the cognitive gap between them.
2. *keywords and topics* can be part of the set P included in the definition of flatten ontology since they have a "global" scope. We can create them by applying a F_{Hybrid} method, as we did with terms. However they are discouraged to be used in isolation, since they focus too much on representativeness and leave aside other details that are essential in the learned ontology.
3. The last alternative leverages on *descriptions or summaries*, which equally have a global nature. The main drawback of such summaries is that they normally focus on describing involved agents

(instances, entities), but ignore more fine-grained ontology details. They can be compared against ground-truth summaries, but more affordable F_{Hybrid} evaluation methods are not feasible since summaries have to be created from scratch.

As we will see in Section 5, most of the ontology learning approaches only consider the first and second dimensions, due to the already-discussed complexity for the third dimension.

5 $M_{OntoLearn}$'s Use Cases in the Literature

Having described our methodology $M_{OntoLearn}$, in this last section we present an analysis of some ontology learning approaches in the literature and the way they have been evaluated, in order to study 1) the way they overlap with our integrated methodology 2) how they can benefit from a better formalised evaluation strategy and the less human-dependent, quality-focused evaluation methods introduced in this paper. The three selected systems are sorted in increasing order of complexity, according to the dimensions in $M_{OntoLearn}$ they target.

Case 1: Learning Domain Ontologies for Web Service Descriptions. In this research work [21] the authors apply their approach over an experimental corpus consisting of 158 EMBOSS bioinformatics service descriptions. Their evaluation looks into the *first dimension* d_W identified in our approach, to measure the lexical precision of the generated ontology by comparing the results against a set of ground truth annotations manually identified from the corpus. This evaluation would have benefited from the application of a $\mathcal{F}_{Hybrid}$ method leveraging on already existing term-spotting techniques in order to reduce the human intervention to a less demanding assessment step.

They tried to tackle also properties (the *relations dimension* d_R), but they found out that de-facto ontologies were too much complex and very different from the extracted ontology. This backs our claim on needing to move from very formal high-level ontologies to more lightweight models. Hence, they put domain experts to rate concepts according to their usefulness in the current task. However, this is not reproducible so other approach working on similar data sets would not be able to reuse such efforts.

Case 2: OntoLearn. OntoLearn [23] evaluation strategy is twofold: first, they provide a detailed quantitative analysis of the ontology learning algorithms, mainly focused on the lexical aspect (d_W). Second, they automatically generate natural language descriptions of formal concept relations in order to facilitate qualitative analysis by domain specialists, therefore targeting d_R.

They claimed that a manual analysis of the extracted terminology is advisable before proceeding with the subsequent steps, arguing that this task lasts about 0.5 minutes per term so it can be easily accomplished in few hours by domain specialists. Those figures back our hypotheses in equation 2b. But unfortunately their assessment is directly made over the results of the algorithm being evaluated, so it cannot be straightforwardly used for developing $\mathcal{F}_{Hybrid}$-based methods that require to leverage on various state-of-the-art automatic algorithms to ensure unbiased evaluations. Concerning the evaluation of relations, they developed a "gloss" generation algorithm in order to facilitate per-concept evaluation by domain specialists. The objetive is to reduce the cost of human assessments, but since they are directly performed over the results they become non reproducible for future campaigns. Other examples of systems considering evaluations fitting into dimensions 1 and 2 are [8] and [17].

Case 3: CRCTOL. The approach presented in [15] has been already introduced in section 2, as one of the most complete evaluations available in the literature. Having now described our methodology $M_{OntoLearn}$, we can study how the considered evaluation objetives fall into our proposed dimensions.

They use the concept "component level" to refer to the lexical aspect expressed by dimension (d_W). They also consider some so-called "relations", including taxonomic and non-taxonomic ones, therefore covering the whole spectrum of connections targeted by dimension d_R. They evaluated these two first dimensions by relying on a manually annotated corpus coming form the US report "Patterns of Global Terrorism Documents", from 1991 to 1994. The problem lays again within the high temporal costs of performing such generation process by relying only on experts in contrast to $\mathcal{F}_{Hybrid}$-like methods. In addition, CRCTOL has been also evaluated in what they named the "Structural Property Based Method". Based on the "small world property" [20] that applies to knowledge networks such as WordNet,

they assume that their automatically built domain ontology should also fit this principle. Hence, they gauge the quality of the built ontology by measuring whether its graph representation is consistent with that of a small world graph. This graph-based evaluation technique has a global scope that allows classifying it into the methods in the third dimension d_P of the methodology.

The organisers of Semeval-2015 [4] also targeted dimension 3 via some structural indicators such as the size of the taxonomy in terms of nodes and edges, the degree of connectivity with the root, and the existence of cycles. Part of their evaluation is very much in line with our $\mathcal{F}_{Hybrid}$ method: taking as input the results submitted by the participants (with affordable $\mathcal{C}(\mathcal{F}_{Auto})$), they asked experts to identify relations which were initially missing in the gold standard (lightweight human intervention $\mathcal{C}(\mathcal{F}_{Assess})$) in order to increase coverage.

6 Conclusions and Future Work

In this position paper we have presented an integrated and incremental methodology $M_{OntoLearn}$ to evaluate the result of ontology learning approaches over big unstructured corpora. Starting from the definition of a *flatten ontology* as a simplified formalisation that better matches the current trends in information extraction and ontology learning, where more specific and changeable domains are involved, we have defined a multidimensional methodology that distributes into three well identified levels of complexity d_W, d_R, d_P the different evaluation objetives that are involved in the learning process: the lexical, the relational, and the global aspects. In addition, we have presented an innovative evaluation method F_{Hybrid} that takes advantage of the reproducibility of corpus-based solutions, while minimising the cost and promoting a higher recall and precision during annotation phase. Through an analysis of previously published efforts on ontology learning systems and their evaluation, we have studied how they fall into some of the evaluation dimensions that we previously identified, and how they could further benefit from applying our methodology' guidelines for a more standardised, less-costly way of targeting the crucial evaluation process.

In the future, we want to keep moving towards less human-based intervention evaluation methods, mainly by further researching on a more automatic and accurate selection of the results coming from the state-of-the-art ontology learner algorithms. Also, we plan to put in practice this methodology in different domains, like the Research Objects considered under the scope of the DrInventor[3] project where we are actively participating, in order to keep confirming and refining our initial claims.

Acknowledgments

This work is partially supported by the FP7 European project Dr Inventor FP7-611383.

References

1. E. Agirre, E. Alfonseca, K. Hall, J. Kravalova, M. Paşca, and A. Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *The 2009 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27, 2009.
2. H. Alani, S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis, and N. R. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, 18(1):14–21, 2003.
3. C. Biemann. Ontology learning from text: A survey of methods. In *LDV forum*, volume 20, pages 75–93, 2005.
4. G. Bordea, P. Buitelaar, S. Faralli, and R. Navigli. Semeval-2015 task 17: Taxonomy extraction evaluation (texeval). *SemEval-2015*, 452(465), 2015.
5. S. L. Camiña. *A comparison of taxonomy generation techniques using bibliometric methods: applied to research strategy formulation.* PhD thesis, Massachusetts Institute of Technology, 2010.
6. P. Cimiano and J. Völker. text2onto. In *International Conference on Application of Natural Language to Information Systems*, pages 227–238, 2005.
7. W. B. Croft, D. Metzler, and T. Strohman. *Search engines: Information retrieval in practice.* Addison-Wesley Reading, 2010.
8. K. Dellschaft and S. Staab. On how to perform a gold standard based evaluation of ontology learning. In *International Semantic Web Conference*, pages 228–241. Springer, 2006.
9. K. Dellschaft and S. Staab. Strategies for the evaluation of ontology learning. In *Ontology Learning and Population*, volume 167, pages 253–272, 2008.
10. M. Ehrig, P. Haase, M. Hefke, and N. Stojanovic. Similarity for ontologies-a comprehensive framework. *ECIS 2005 Proceedings*, page 127, 2005.
11. X. Gao, B. Xiao, D. Tao, and X. Li. A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129, 2010.

[3] `http://drinventor.eu/`

12. J. L. R. García, G. Rizzo, and R. Troncy. The Concentric Nature of News Semantic Snapshots. In *8th international Conference on Knowledge Capture (KCAP)*, 2015.
13. N. Guarino and C. A. Welty. An overview of ontoclean. In *Handbook on ontologies*, pages 201–220. Springer, 2009.
14. U. Hahn and M. Romacker. The syndikate text knowledge base generator. In *1st international conference on Human language technology research*, pages 1–6, 2001.
15. X. Jiang and A.-H. Tan. Crctol: A semantic-based domain ontology learning system. *Journal of the American Society for Information Science and Technology*, 61(1):150–168, 2010.
16. J. P. Leal, V. Rodrigues, and R. Queirós. Computing semantic relatedness using dbpedia. In *OASIcs-OpenAccess Series in Informatics*, volume 21, 2012.
17. D. Manzano, A. Gómez-Pérez, and D. Borrajo. Unsupervised and domain independent ontology learning: combining heterogeneous sources of evidence. 2008.
18. J. Palmius. Criteria for measuring and comparing information systems. 2007.
19. R. Porzel and R. Malaka. A task-based approach for ontology evaluation. In *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. Citeseer, 2004.
20. J. Ramanand, A. Ukey, B. K. Singh, and P. Bhattacharyya. Mapping and structural analysis of multi-lingual wordnets. *IEEE Data Eng. Bull.*, 30(1):30–43, 2007.
21. M. Sabou, C. Wroe, C. Goble, and G. Mishne. Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In *14th international conference on World Wide Web*, pages 190–198. ACM, 2005.
22. H. Tanev and B. Magnini. Weakly supervised approaches for ontology population. In *Conference on Ontology Learning and Population*, pages 129–143, 2008.
23. P. Velardi, R. Navigli, A. Cuchiarelli, and R. Neri. Evaluation of ontolearn, a methodology for automatic learning of domain ontologies. *Ontology Learning from Text: Methods, evaluation and applications*, 123:92, 2005.
24. W. Wong, W. Liu, and M. Bennamoun. Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, 44(4):20, 2012.
25. L. A. Zager and G. C. Verghese. Graph similarity scoring and matching. *Applied mathematics letters*, 21(1):86–94, 2008.
26. E. Zavitsanos, G. Paliouras, and G. A. Vouros. Gold standard evaluation of ontology learning methods through ontology transformation and alignment. *IEEE Transactions on Knowledge and Data Engineering*, 23(11):1635–1648, 2011.

Bibliografía

[Berland and Charniak, 1999] Berland, M. and Charniak, E. (1999). Finding parts in very large corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 57–64. Association for Computational Linguistics.

[Brank et al., 2005] Brank, J., Grobelnik, M., and Mladenic, D. (2005). A survey of ontology evaluation techniques. In *Proceedings of the conference on data mining and data warehouses (SiKDD 2005)*, pages 166–170.

[Brewster et al., 2004] Brewster, C., Alani, H., Dasmahapatra, S., and Wilks, Y. (2004). Data driven ontology evaluation.

[Brill, 1992] Brill, E. (1992). A simple rule-based part of speech tagger. In *Proceedings of the workshop on Speech and Natural Language*, pages 112–116. Association for Computational Linguistics.

[Chaves, 2016] Chaves, D. (2016). Dataset of Terms from Computer Graphics Doamin. `https://dx.doi.org/10.6084/m9.figshare.3485690.v2`.

[Cimiano et al., 2004] Cimiano, P., Schmidt-Thieme, L., Pivk, A., and Staab, S. (2004). Learning taxonomic relations from heterogeneous evidence. In *Proceedings of the ECAI 2004 Ontology Learning and Population Workshop*.

[Daelemans and Reinberger, 2004] Daelemans, W. and Reinberger, M. (2004). Shallow text understanding for ontology content evaluation. *IEEE Intelligent Systems*, 19(4):74–81.

[Dellschaft and Staab, 2008] Dellschaft, K. and Staab, S. (2008). Strategies for the evaluation of ontology learning. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, Frontiers in Artificial Intelligence and Applications*, volume 167, pages 253–272. Citeseer.

[Fernández-López et al., 1997] Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: from ontological art towards ontological engineering.

[Fleiss and Cohen, 1973] Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement.*

[Frantzi et al., 2000] Frantzi, K., Ananiadou, S., and Mima, H. (2000). Automatic recognition of multi-word terms:. the c-value/nc-value method. *International Journal on Digital Libraries*, 3(2):115–130.

[Galton, 1892] Galton, F. (1892). *Finger prints.* Macmillan and Company.

[Gangemi et al., 2006] Gangemi, A., Catenacci, C., Ciaramita, M., and Lehmann, J. (2006). *Modelling ontology evaluation and validation.* Springer.

[Girju et al., 2002] Girju, R., Moldovan, D. I., et al. (2002). Text mining for causal relations. In *FLAIRS Conference*, pages 360–364.

[Gómez-Pérez, 1999] Gómez-Pérez, A. (1999). Evaluation of taxonomic knowledge in ontologies and knowledge bases.

[Gómez-Pérez, 2004] Gómez-Pérez, A. (2004). Ontology evaluation. In *Handbook on ontologies*, pages 251–273. Springer.

[Guarino, 1998] Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. *arXiv preprint cmp-lg/9809002.*

[Guarino and Welty, 2009] Guarino, N. and Welty, C. A. (2009). An overview of ontoclean. In *Handbook on ontologies*, pages 201–220. Springer.

[Hahn and Schnattinger, 1998] Hahn, U. and Schnattinger, K. (1998). Towards text knowledge engineering. *Hypothesis*, 1(2).

[Lavrac and Dzeroski, 1994] Lavrac, N. and Dzeroski, S. (1994). Inductive logic programming. In *WLP*, pages 146–160. Springer.

[Maedche, 2012] Maedche, A. (2012). *Ontology learning for the semantic web*, volume 665. Springer Science & Business Media.

[Maedche and Staab, 2002] Maedche, A. and Staab, S. (2002). Measuring similarity between ontologies. In *Knowledge engineering and knowledge management: Ontologies and the semantic web*, pages 251–263. Springer.

[Maedche and Volz, 2001] Maedche, A. and Volz, R. (2001). The ontology extraction & maintenance framework text-to-onto. In *Proc. Workshop on Integrating Data Mining and Knowledge Management, USA*, pages 1–12.

[Manzano-Macho et al., 2008] Manzano-Macho, D., Gómez-Pérez, A., and Borrajo Millán, D. (2008). Unsupervised and domain independent ontology learning: combining heterogeneous sources of evidence.

[Maynard et al., 2008] Maynard, D., Li, Y., and Peters, W. (2008). Nlp techniques for term extraction and ontology population. In *Proceeding of the 2008 conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, pages 107–127.

[Maynard et al., 2006] Maynard, D., Peters, W., and Li, Y. (2006). Metrics for evaluation of ontology-based information extraction. In *International world wide web conference*, pages 1–8. Edinburgh, UK.

[Pinto and Martins, 2004] Pinto, H. S. and Martins, J. P. (2004). Ontologies: how can they be built? *Knowledge and information systems*, 6(4):441–464.

[Porzel and Malaka, 2004] Porzel, R. and Malaka, R. (2004). A task-based approach for ontology evaluation. In *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*. Citeseer.

[Sabou et al., 2005] Sabou, M., Wroe, C., Goble, C., and Mishne, G. (2005). Learning domain ontologies for web service descriptions: an experiment in bioinformatics. In *Proceedings of the 14th international conference on World Wide Web*, pages 190–198. ACM.

[Schmid, 2013] Schmid, H. (2013). Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154. Routledge.

[Shamsfard and Barforoush, 2004] Shamsfard, M. and Barforoush, A. A. (2004). Learning ontologies from natural language texts. *International journal of human-computer studies*, 60(1):17–63.

[Spyns, 2005] Spyns, P. (2005). Evalexon: Assessing triples m ined from texts. *STAR*, 2005(09):09.

[Spyns and Reinberger, 2005] Spyns, P. and Reinberger, M.-L. (2005). Lexically evaluating ontology triples generated automatically from texts. In *The semantic web: Research and applications*, pages 563–577. Springer.

[Studer et al., 1998] Studer, R., Benjamins, V. R., and Fensel, D. (1998). Knowledge engineering: principles and methods. *Data & knowledge engineering*, 25(1):161–197.

[Turney, 2001] Turney, P. (2001). Mining the web for synonyms: Pmi-ir versus lsa on toefl.

[Velardi et al., 2005] Velardi, P., Navigli, R., Cuchiarelli, A., and Neri, R. (2005). Evaluation of ontolearn, a methodology for automatic learning of domain ontologies. *Ontology Learning from Text: Methods, evaluation and applications*, 123:92.

[Wong et al., 2012] Wong, W., Liu, W., and Bennamoun, M. (2012). Ontology learning from text: A look back and into the future. *ACM Computing Surveys (CSUR)*, 44(4):20.